TITULNÍ LIST PERIODICKÉ ZPRÁVY 2006 PROJEKTU 2C06009

Ministerstvo školství, mládeže a tělovýchovy

2C06009

PROSTŘEDKY TVORBY KOMPLEXNÍ BÁZE ZNALOSTÍ PRO KOMUNIKACI SE SÉMANTICKÝM WEBEM V PŘIROZENÉM JAZYCE

řešitel - **doc. Ing. Karel Ježek, CSc.**

.....

(podpis)

za příjemce - **Západočeská univerzita v Plzni** (IČ: 49777513)

rektor

Doc. Ing. Josef Průša, CSc.

.....

(podpis, razítko)

Verze zprávy: **1** Zpracováno dne: **25.1.2007**

2. SKUTEČNOST ZA UPLYNULÉ OBDOBÍ - 2006

2.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

2.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 8/
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

2.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Albrecht Štěpán Ing.** 810520/2061 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 496 377 632 402 albrs@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Fiala Dalibor Ing.** 8003235845 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 dalfa@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Hanks Patrick Ph.D.** 400324 GB
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	hanks@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Horák Aleš Ph.D.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Hynek Jiří ing. PhD** 720506/2029 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632455 hynekj@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky

Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	40

Celé jméno, RČ	**Král Pavel ing.** 760317/2049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky

Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pavelka Tomáš Ing.** 7909182083 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd katedra informatiky a výpočetní techniky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Ptáčková Helena** 705914/2079 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	5

Celé jméno, RČ	**Rychlý Pavel Mgr. Ph.D.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace

Pracovní kapacita v %	50

Celé jméno, RČ	**Sojka Petr RNDr. Ph.D.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Steinberger Josef Ing.** 7909182127 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 479 jstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Tesař Roman Ing.** 7909302379 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 romant@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Toman Michal Ing.** 8007042054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 mtoman@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

2.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2006

Pč.	Typ	Popis
1	změny v projektovém týmu a řešitelských týmech	Dva z plánovaných pracovníků (řádní doktorandi D. Andrš a V. Beneš, kterým končilo doktorandské studium k 31.8.2006) nenastoupili do pracovního poměru na ZČU a odešli z univerzity do praxe. Pro řešení plánovaných prací byli za ně přijati jiní pracovníci (Š. Albrecht, P.Král, J.Hynek), avšak s jistou časovou prodlevou nutnou pro výběrové a administrativní úkony. Jeden z členů kolektivu (D. Fiala), který je v doktorandském studiu pod dvojím vedením, byl v říjnu povolán francouzským spoluškolitelem na univerzitu ve Štrasburku. I když ve Francii pracuje na úkolech souvisejících s výzkumem našeho projektu, neměli jsme dostatek prostředků na jeho vyslání formou dlouhodobé zahraniční pracovní cesty. Požádal proto na doporučení právního odboru univerzity o neplacené pracovní volno na dobu svého pobytu ve Francii. Po návratu ze studijního pobytu v roce 2007 bude opětovně zařazen do řešitelského týmu projektu. Pro administrativní úkony byla do projektu zapojena Helena Ptáčková.

2.2. ČASOVÝ POSTUP PRACÍ

Komentář k metodice a časovému postupu prací a průběhu aktivit za uplynulé období

V roce 2006 bylo zahájeno řešení projektu souborem přípravných prací, jejichž cílem je vytvoření nezbytných metodických a datových souborů, s jejichž pomocí bude plánovaného výsledku projektu úspěšně dosaženo. Mezi nejdůležitější složky patří vytváření různých korpusů, ať již textových nebo řečových, bez nichž by dosažení plánovaného cíle nebylo možné. Detailnější popis jednotlivých vytvářených korpusů je možno nalézt v podrobném popisu jednotlivých aktivit, je však třeba poznamenat, že vytváření obsáhlých korpusů je dlouhodobá záležitost, první výsledky mohou být k dispozici koncem roku 2007, avšak s tím, že korpusy budou po celou dobu řešení projektu nadále doplňovány. Detaily jsou uvedeny v jednotlivých položkách odstavce 2.2.1.
Veškeré aktivity plánované na druhé pololetí 2006 se podařilo splnit, dílčí výsledky dosažené v jednotlivých aktivitách jsou uvedeny v odst. 2.2.1 a v přílohách (odst. 4.1.1 a 4.2.1).

2.2.1. AKTIVITY USKUTEČNĚNÉ v roce 2006

Číslo aktivity

01

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava pořizování korpusu LAC-SS2006

Zahájení aktivity

10.7.2006

Ukončení aktivity

13.10.2006

Popis aktivity

V rámci aktivity 2006-01 byla činnost dílčího týmu zaměřena na přípravu potřebné infrastruktury k pořízení korpusu spontánní řeči LAC-SS2006. Nejprve byla pozornost zaměřena na analýzu charakteru záznamů v pořizovaném korpusu, neboť ten významně ovlivňuje mj. výběr transkripčního software, transkripčních značek, apod. Provedeno bylo několik zkušebních záznamů, aby bylo možno posoudit míru výskytu mimojazykových prvků, ruchů, hovorových výrazů, cizojazyčných a případně jinak obtížně transkribovatelných výrazů, anakolutů, atp. Na základě poslechu těchto zkušebních záznamů byla pak zvolena metodu transkripce korpusu. Řešitelé dospěli k rozhodnutí provádět transkripci ortografickou s tím, že později bude ortografický přepis automaticky konvertován do tvaru fonetického. Dále bylo v rámci aktivity 2006-01 otestováno několik (pět různých) volně dostupných transkripčních programů. Po poměrně rozsáhlém testování byl vybrán byl program Transcriber 1.5.0, který je k dispozici v rámci licence GPL. V přípravné fázi byl také proveden pečlivý výběr technického vybavení, a to jednak k pořizování záznamů spontánní řeči a jednak pro její transkripci. Detaily jsou uvedeny v odstavci 4.1.

Skutečné Indikátory dosažení - výsledky aktivity

1) volba ortografické transkripce s tím, že později bude ortografický přepis automaticky konvertován do tvaru fonetického,

2) návrh souboru transkripčních značek,

3) volba transkripčního programu Transcriber 1.5.0,

4) výběr a pořízení technického vybavení, a to jednak k pořizování záznamů spontánní řeči a jednak pro její transkripci.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Jednalo se o přípravnou fázi, na jejíž důkladném provedení závisí mnoho dalších aktivit plánovaných při řešení projektu. Výběr, návrhy a nákup byly provedeny velmi obezřetně, aby řečová data korpusu mohla být pořízena kvalitně a ukládána bez nadbytečné informace.

Číslo aktivity

02

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření software pro editaci sémantických anotací

Zahájení aktivity

3.7.2006

Ukončení aktivity

30.11.2006

Popis aktivity

Tvorba sémantických anotací vět přirozeného jazyka je časově i finančně náročná. Vhodný software, který umožní

anotátorům snazší a rychlejší sémantickou anotaci, je klíčový pro snížení nákladů nutných pro sémantické označkování korpusu. Potřebný software byl vyvíjen s požadavkem maximální uživatelské přívětivosti a robustnosti. Anotatátor je při anotaci veden anotačními schématy, která popisují, jakým způsobem je možné větu anotovat. Anotační schémata jsou uložena v XML formátu, k formátu existuje XML schéma a soubor je vždy podle XML schématu validován. Vytvořený software má následující vlastnosti: o načítání vstupních vět pro anotaci, načítání souborů s již anotovanými větami o přívětivý GUI o podpora více anotačních schémat o verzování anotačních schémat o periodické automatické ukládání (pro zálohu) o multiplatformní běh

Skutečné Indikátory dosažení - výsledky aktivity

Komplexní softwarové vybavení pro podporu ručního sémantického značkování korpusu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Bylo prováděno testování lidmi přijatými na sémantické značkování korpusu. Seznam problémů a návrhů byl do programu postupně zapracován. Testování současné verze programu (verze 1.3) neprokázalo žádné významné nedostatky, které by bránily přímému nasazení v anotačním procesu.

Číslo aktivity

03

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Výběr a zaškolení pracovníků provádějících transkripci

Zahájení aktivity

16.10.2006

Ukončení aktivity

24.11.2006

Popis aktivity

Prostřednictvím veřejného oznámení jsme nabídli spolupráci (provádění transkripce záznamů v korpusu LAC-SS2006) studentům Filozofické fakulty se zaměřením na jazyky. Ze studentů, kteří měli o spolupráci zájem, jsme vybrali dvě studentky, které v testu (náplní testu bylo ověření schopností a vloh pro provádění transkripce přepisem reálných záznamů) uspěly nejlépe. Tyto spolupracovnice byly následně vyškoleny částečně v oblasti zpracování zvukového záznamu, zejména pak v oblasti ortografické a fonetické transkripce a ovládání transkripčního software.

Skutečné Indikátory dosažení - výsledky aktivity

V současné době jsou plně vyškoleny dvě studentky, které provádějí transkripci záznamů, další student je momentálně školen a jeho aktivita bude využita v roce 2007.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vybrány a vyškoleny byly studentky filozofické fakulty:
Marie Škochová – F04126
Gabriela Wagnerová – F04148

Číslo aktivity

04

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava a vytváření korpusu vět z vybraných domén

Zahájení aktivity

4.9.2006

Ukončení aktivity

Popis aktivity

Studentům fakulty aplikovaných věd jsme jako součást semestrální práce zadali vytvoření textového souboru obsahujícího průměrně 300 záznamů vět (počet vět kolísal dle studijního zaměření studenta) z deseti různých oblastí (domén). Studenti pracovali ve dvojicích, celkově se projektu zúčastnilo 180 studentů. Studenti obdrželi manuál, jakým způsobem mají záznamy vytvářet, nejdůležitější bylo vytvářet přirozené věty a v případě nutnosti požádat další účastníky (rodiče, přátele apod.) o spolupráci. Studenti vytvářeli věty v následujících doménách: o počasí (např. Bude zítra pršet v Plzni?) o nakupování (např. Kde bych sehnal nejlevnějšího eriksona? o městská doprava (Kdy mi to jede po desátý na Košutku?) o ubytování (Kolik stojí noc v kontíku?) o restaurace (Kde se tady dá rozumně najíst?) o městské a státní úřady (Od kolika ráno maj na finančáku?) o památky, muzea (Je přes zimu otevřený plzeňský podzemí?) o vlakové a autobusové spoje (Potřebovala bych dojet do devíti do Klatov) o 2 domény volné (při výběru však ověřte, že existuje vhodná databáze problémové oblasti na webu) První kontrola výsledků se uskutečnila v týdnu 6.11. – 10.11., studenti měli vytvořeny asi dvě třetiny požadovaných vět, tj. celkem cca 18000 vět (90 dvojic*200vět). Dle jejich údajů se na tvorbě každého korpusu 200 vět podílelo průměrně 5 lidí. Celkem tedy můžeme očekávat korpus od 450 respondentů. Cca 70% vět je použitelných pro další zpracování (sémantickou anotaci), některé z vět již byly pro sémantickou anotaci použity. V období 27.11. – 21.12. odevzdají studenti výsledky své práce; očekáváme vytvoření celkem cca 27000 vět, z toho cca 19000 vět využitelných pro další práci.

Skutečné Indikátory dosažení - výsledky aktivity

V současné době je k dispozici 18000 vět, očekáváme vytvoření celkem 27000 vět do konce ledna 2007.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Je prováděna ruční kontrola souborů korpusu + testování lidmi přijatými pro sémantické značkování korpusu. Jelikož práce dosud nejsou ukončeny, hodnocení zatím neuvádíme.

Číslo aktivity

05

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Pořizování záznamů spontánní řeči, transkripce a její supervize (fáze 1)

Zahájení aktivity

30.10.2006

Ukončení aktivity

Popis aktivity

Nejprve se pracovníci, kteří se v rámci své odbornosti nezabývají přímo zpracováním digitálního signálu, seznámili s nástroji pro zpracování digitálního signálu tak, aby byli schopni upravovat pořízené záznamy (zejména stříhat, upravovat celkový dynamický rozsah, apod.). Byli částečně proškoleni v oblasti elektroakustiky a mikrofonní techniky, aby se nedopouštěli chyb při pořizování záznamů do korpusu. Poté jsme pravidelně pořizovali nahrávky spontánních řečových projevů (v 1. fázi zejména přednášek pracovníků laboratoře, seminářů studentů doktorandského studia, seminářů diplomantů a prezentací studentů předmětů, jejichž výuka je zajišťována pracovníky laboratoře). Tyto nahrávky byly následně upraveny tak, aby byly vhodné k transkripci. Upravené záznamy následně vyškolení pracovníci transkribovali pomocí nástroje Transcriber 1.5.0 podle pokynů uvedených v transkripčním manuálu. Pracovníci laboratoře v průběhu prvních dvou týdnů kontrolovali kvalitu transkripce vytvořené pracovníky provádějící transkripci, poskytovali jim rady (zejména ve sporných a nejednoznačných případech) a zároveň sbírali data a poznatky k obohacení a případným úpravám transkripčního manuálu. V této

fázi jsme např. zjistili, že je třeba zavést některé další transkripční značky a postupy transkripce v případech, které nebyly v době prvotního návrhu mechanismů transkripce známy. Na konci této 1. fáze jsme provedli ověření (úspěšné), zda jsou spolupracovníci provádějící transkripci schopni pracovat zcela samostatně a řešit případné nastalé problémy bez asistence pracovníka laboratoře.

Skutečné Indikátory dosažení - výsledky aktivity

V současné době je pořízeno 12,38 hodin nahrávek spontánní řeči, z toho 2,31 hodin je kvalitně transkribováno.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Záznamy spontánní řeči jsou nadále pořizovány a zpracovávány, zpracování a ověření výsledků bude provedeno v následujícím kalendářním roce.

Číslo aktivity

06

Ke kterému dílčímu cíli se aktivita vztahuje

Název (cíl)aktivity

Vytvoření prezentačních webových stránek projektu

Zahájení aktivity

20.8.2006

Ukončení aktivity

1.12.2006

Popis aktivity

Byl upraven výkonný redakční systém Mediawiki dostupný v rámci licence GPL tak, aby fungoval jako inteligentní studnice informací, pomocí níž si mohou jednotliví členové řešitelského týmu vyměňovat zprávy, data a poznatky a který umožňuje některé oblasti této informační studnice zpřístupnit veřejnosti jako webové stránky. Pro tento systém jsme navrhli moderní reprezentativní grafiku, která odráží charakter laboratoře i řešené problematiky. Instalovali jsme výkonný a spolehlivý 64-bitový stroj Sun pro funkci webového serveru, bezpečný operační systém Ubuntu server (Linux 2.6.17) a vše potřebné k provozu webového portálu. Zároveň jsme instalovali servletový kontejner Apache Tomcat 5.5, díky kterému budeme moci webový portál rozšiřovat i o vysoce interaktivní služby založené na technologii Java Servlet.

Skutečné Indikátory dosažení - výsledky aktivity

Výkonný portál založený na redakčním systému Mediawiki, moderní design webových stránek.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Viz http://liks.fav.zcu.cz/mediawiki/index.php/Research#COT-SEWing

---

Číslo aktivity

07

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Výběr a zaškolení pracovníků provádějících sémantické anotace

Zahájení aktivity

1.10.2006

Ukončení aktivity

30.11.2006

Popis aktivity

Cílem této akce bylo získat 4 kvalifikované pracovníky pro sémantické značkování korpusu. Formou veřejného oznámení byla nabídnuta spolupráce studentům Filosofické fakulty se zaměřením na jazyky. Do konkurzu se přihlásilo celkem 7 lidí. Samotný konkurz probíhal od 13. 11. do 16. 11. 2006. Během konkurzu byla testována zejména schopnost abstraktního myšlení. Kandidáti prováděli čtyři úkoly, které budou jejich pracovní náplní

(anotační manuál a popis úkolů dostali tištěný k dispozici – viz http://liks.fav.zcu.cz/mediawiki/images/a/a8/AnManKon.pdf). 16.11. 2006 byl konkurz uzavřen a výsledky bodově vyhodnoceny. Takto získaný tým spolupracovníků byl vyškolen v oblasti sémantických anotací. Splnění tohoto úkolu vyžadovalo částečně splněný cíl aktivity 2006-02 (software) a cíl aktivity 2006-04 (data).

Skutečné Indikátory dosažení - výsledky aktivity

Čtyři plně vyškolení pracovníci, kteří budou provádět sémantické anotace:
Petra Hajdúchová – K04638
Hana Heřmanová – R05097
Michaela Matějovicová – F04088
Martin Moravec – F05099

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Během anotací bude zkoumána shoda mezi anotátory. Míra shody bude vypovídat jak o přesnosti navrženého sémantického popisu, tak o úrovni techniky anotátorů.

Číslo aktivity

08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

LGMM – LASER Gaussian Mixture Module v. 1.0

Zahájení aktivity

3.7.2006

Ukončení aktivity

31.10.2006

Popis aktivity

Vytvoření modulu ASR systému LASER, který převádí parametrizovaný řečový signál na emisní pravděpodobnosti stavů Markovova modelu (HMM). Parametry HMM jsou trénovány toolkitem HTK. Program musí řešit numerickou stabilitu při použití směsi vícerozměrných Gaussových funkcí. Současná verze nepodporuje tzv. "svazování" stavů, které se používá při použití kontextově závislých fonetických jednotek.

Skutečné Indikátory dosažení - výsledky aktivity

Vytvoření části funkčního softwaru, který bude používán v dalších fázích řešení projektu.

Publikace v: Pavelka, T.: LDec: One Pass Time Synchronous Decoder, PhD. Workshop 2006, Hrubá Skála, Czech Republic

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vytvořený software je integrální součástí vytvářeného softwarového systému pro hlasovou komunikaci se sémantickým webem, systém bude nadále rozšiřován o další programové moduly.

Číslo aktivity

09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Detektor ticha a řeči (1)

Zahájení aktivity

4.9.2006

Ukončení aktivity

Popis aktivity

Návrh formálního modelu pro detekci ticha a řeči založeného na GMM/HMM a jeho testování – první fáze návrhu.

Skutečné Indikátory dosažení - výsledky aktivity

Sada skriptových souborů s detekčním a testovacím software, součást budovaného programového systému.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byla vytvořena statistika chyb.

Číslo aktivity

10

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vytvoření software pro převod EBNF gramatik na HMM

Zahájení aktivity

1.9.2006

Ukončení aktivity

15.12.2006

Popis aktivity

Vytvoření programu pro převod gramatiky ve formátu EBNF (Extended Backus-Naur Form) a slovníku s fonetickou výslovností na Markovův model (HMM), který definuje všechny rozpoznávané promluvy. Částečně řeší minimalizaci stavů modelu.

Skutečné Indikátory dosažení - výsledky aktivity

Vytvoření modulu funkčního softwaru (součást budovaného programového systému), který bude využíván v dalších fázích řešení projektu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Součást budovaného programového systému.

Číslo aktivity

11

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Výběr vhodného typu neuronové sítě pro zpracování sémantiky přirozeného jazyka

Zahájení aktivity

10.7.2006

Ukončení aktivity

30.11.2006

Popis aktivity

V rámci dostupných popisů a implementací neuronových sítí byla hledána aplikovatelnost použití neuronových sítí v oblasti zpracování sémantiky jazyka a poté nejvhodnější neuronová síť pro zpracování vstupů přirozeného jazyka. Členové týmu procházeli desítky článků a webových stránek výzkumných center, prostudovali souvislosti mezi biologickými sítěmi zpracovávající sémantiku přirozeného jazyka (lidký mozek) a umělými neuronovými sítěmi. Jako možnou aplikovatelnou pak vyhodnotili Kohonenovu samoorganizující mapu (SOM). Zkoumány byly nutné podmínky pro vstupní vektory této mapy v souvislosti ze zpracováním přirozeného jazyka (WEBSOM).

Skutečné Indikátory dosažení - výsledky aktivity

Na základě prováděných testů a dalších posouzení byla vybrána neuronová síť WEBSOM.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Zatím pouze na základě dostupných pramenů, připravuje se testování na vytvářeném korpusu.

Číslo aktivity

12

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Ověření vlastností vhodného typu neuronové sítě pro zpracování sémantiky přirozeného jazyka

Zahájení aktivity

1.11.2006

Ukončení aktivity

Popis aktivity

V prvé fázi řešení projektu byla ověřována možnost využití některých umělých neuronových sítí pro automatické získávání informací z kolekcí dokumentů a pro organizování velkých kolekcí na základě podobnosti textů. Z dostupných publikací (viz předcházející bod) zabývajících se touto problematikou je patrné, že jako perspektivní pro tuto oblast se jeví Kohonenova samoorganizující mapa (SOM), jejíž modifikace (WEBSOM) byla využita k vytváření shluků dokumentů s podobným obsahem a k následnému vyhledávání v těchto dokumentech. V další fázi řešení projektu se tedy zaměříme převážně na této neuronovou síť. V současnosti probíhá testování volně dostupných (ale zároveň kvalitních) simulátorů této neuronové sítě (SOM_PAK a SOM toolbox) za účelem ověření, zda bude možné jejich další využití, popř. zda bude nutné provést vlastní implementaci. Zmíněná metoda WEBSOM byla autory ověřena na dokumentech psaných v angličtině a finštině. Tyto dokumenty byly za účelem testování podobnosti kódovány s využitím tzv. mapy slovních spojení, založené na využití kontextové informace obsažené v dokumentech. Naším dalším úkolem tedy bude ověření možnosti využití Kohonenovy mapy pro zpracování kolekcí česky psaných dokumentů a zároveň vytvoření vhodného popisu (popř. vyžití popisu použitého v na základním algoritmu WSOM), na základě kterého by mohla být u česky psaných dokumentů posuzována jejich vzájemná podobnost.

Skutečné Indikátory dosažení - výsledky aktivity

Byly nainstalovány volně dostupné simulátory Kohonenovy mapy (SOMPAK a SOMtoolbox) za účelem ověření funkčnosti a vhodnosti Kohonenovy mapy pro získávání informací z kolekcí dokumentů a pro organizování velkých kolekcí česky psaných dokumentů. Tyto simulátory jsou funkční a počáteční testy ukázaly, že po doplnění vhodného rozhraní by mohly být využitelné v prvé fázi řešení projektu, pro účely zjišťování podobnosti dokumentů psaných v češtině. V konečné fázi, pro začlenění Kohonenovy mapy do výsledného systému, pak bude pravděpodobně nutná vlastní implementace.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

V prvé fázi se jednalo pouze o testování dostupných simulátorů, výsledky testů posuzování podobnosti kolekcí česky psaných dokumentů bude provedeno později.

Číslo aktivity

13

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

ARAD modul – modul pro automatické rozpoznávání dialogových aktů

Zahájení aktivity

4.7.2006

Ukončení aktivity

15.12.2006

Popis aktivity

Automatické rozpoznávání aktů dialogu (angl. dialogue acts), jako je sdělení, otázka, souhlas, apod., je jedním ze základních prvků porozumění spontánního dialogu. V rámci této aktivity jsme se proto zaměřili na návrh a implementaci modulu, který umožní automaticky rozpoznat akty dialogu v rozhovoru dvou či více osob. Navržený modul se skládá ze dvou částí. První část využívá lexikální (a syntaktickou) informaci, která je uložena ve formě rozpoznaných slov (výstup recognizeru). Druhá část pracuje s prozodickou informací, kterou představují vypočtené atributy základní hlasivkové frekvence (F0) a energie v řečovém signálu. Klasifikace rozhovoru do tříd aktů dialogu probíhá paralelně v obou částech analýzy (lexikální a prozodická analýza) modulu, přičemž výsledek rozpoznání je kombinací obou dílčích výsledků. Lexikální část využívá pro rozpoznávání moderní statistické metody, jako jsou n-gramy, apod. Prozodická část používá k rozpoznávání model směsi Gaussových funkcí (angl. Gaussian Mixture model). Ke kombinaci obou dílčích výsledků je zatím použita neuronová síť typu MLP.

Skutečné Indikátory dosažení - výsledky aktivity

Jednoduchý funkční prototyp, možnost jeho použití v dalších fázích projektu.

Publikace:

Kral, P., Kleckova, J., and Cerisasa, C.: Automatic Dialog Acts Recognition based on Words Clusters. In WESPAC IX 2006, Seoul, Korea, 2006.

Kral, P., Cerisasa, C., and Kleckova, J.: Automatic Dialog Acts Recognition based on Sentence Structure. In: ICASSP '06 Proceedings, Toulouse, France, 2006, pp. 61-64.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Modul byl otestován na čtyřech základních aktech dialogu (sdělení, otázka zjišťovací, otázka doplňovací a příkaz) obsažených v testovacím korpusu dialogů zaznamenaných při dotazech na možná autobusová a železniční spojení.

Číslo aktivity

14

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem

Zahájení aktivity

15.10.2006

Ukončení aktivity

Popis aktivity

Pro automatické rozpoznávání aktů dialogu bylo potřeba nadefinovat množinu aktů dialogu spolu s vhodným systémem značkování korpusu. V rámci této aktivity jsme se proto zaměřili na definování množiny aktů dialogu a vhodného systému označkování korpusu. Bylo potřeba vyřešit následující dva problémy: 1) množina aktů dialogu musí být dostatečně obecná, aby byla jednoduše použitelná i v rámci jiných úloh, 2) definice aktů dialogu musí být dostatečně jednoduchá, aby značkování korpusu bylo rychlé a nedocházelo k nejednoznačnostem při značkování. Nově nadefinovaná množina aktů dialogu je založena na populárním projektu DAMSL (Dialogue Act Markup in Several Layers) a MRDA (Meeting Recorder of Dialogue Acts) taxonomii. DAMSL definuje pokud možno univerzální množinu aktů dialogu (42 tříd) a jeho anotační schéma je složeno ze čtyř úrovní: komunikační status, informační rovina, dopředovazební funkce a zpětnovazební funkce. Množina aktů dialogu v projektu MRDA vychází z taxonomie SWBD-DAMSL (drobně upravená množina DAMSL). Značky aktů dialogu zde nejsou organizovány podle výše uvedených čtyř úrovní, ale každý akt dialogu je charakterizován obecnou značkou a v případě nejednoznačnosti je přidána značka specifická.

Skutečné Indikátory dosažení - výsledky aktivity

Byla navržena množina aktů dialogu a systém jejich značkování.

Publikace:
Kral, P., Cerisara, C., Kleckova, J., Pavelka, T.: Sentence Structure for Dialog Act Recognition in Czech. Proceedings of ICTTA´06, Damascus, Syria, 2006

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Jedná se pouze o dílčí úkol, jehož výsledky budou využity v dalším období pro návrh systému řízení dialogu, dialog-manageru atd.

Číslo aktivity
15
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Vytváření vícejazyčných korpusů
Zahájení aktivity
3.7.2006
Ukončení aktivity

Popis aktivity
Pro ověřování vlastností vyvíjeného programového vybavení je nutno vytvořit korpusy zaměřené na konkrétní problémové oblasti. Na počátku byla pozornost soustředěna na korpusy obsahující texty omezené domény počítačových kateder. V rámci aktivity byl vytvořen korpus českých textů shrnutím dokumentů nalezených na 17 serverech českých kateder informatiky v létě 2006 (zatím má rozsah cca 7 GB) a pak korpus francouzských dokumentů o rozsahu téměř 40 GB, který obsahuje převzaté dokumenty z 80 francouzských univerzitních počítačových laboratoří.
Skutečné Indikátory dosažení - výsledky aktivity
Vytvářené korpusy budou sloužit především pro zpracování textových informací a budou po celou dobu řešení projektu průběžně doplňovány v závislosti na problémových oblastech, které budou v rámci řešení projektu zpracovávány.

Publikace:
Fiala D., Tesař R., Ježek K., Rousselot F.: “Extracting Information from Web Content and Structure”. The 9th International Conference on Information Systems Implemen-tation and Modelling (ISIM ’06), Přerov, Czech Republic, ISBN 80-86840-19-0, pages 133-140, CEUR-WS proceedings, Vol. 180, ISSN 1613-0073, http://ceur-ws.org/Vol-180, 2006.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Zatím pouze vizuální a po dohodě – vzhledem ke značným velikostem není možné korpusy umístit na web, ale je nutno způsob jejich získání domluvit emailem na adresu dalfia@kiv.zcu.cz.

Číslo aktivity
16
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Koncepce programového vybavení pro vytváření vícejazyčných korpusů
Zahájení aktivity

25.8.2006

Ukončení aktivity

30.11.2006

Popis aktivity

V rámci příprav na vytváření a zpracování korpusových dat byl vytvořen programový subsystém umožňující pohodlné pořizování a následné paměťově nenáročné ukládání korpusových dat do paměti. Programové řešení bylo zpracování jako relativně autonomní programový subsystém zpracovaný v jazyce C# a sestává z celkem pěti programových modulů: modul WebWatch3 vytváří korpus webových dokumentů, modul Convertor převádí dokumenty v nejrůznějších formátech do textu, modul Analyzer analyzuje dokumenty a statistickými metodami z nich extrahuje informace nutné pro vytvoření citační sítě autorů, modul Aggregator připravuje databázi autorů k následujícím výpočtům autoritativnosti, modul RankCalculator implementuje několik známých a několik nových algoritmů pro zjišťování významnosti uzlů grafu (použito na získaný webový graf stránek akademických institucí a na citační graf odborných publikací).

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem této aktivity je vytvoření nezávislého programového vybavení pro vytváření datových korpusů, které bude sloužit především pro zpracování textových informací. Hotové programové řešení představuje první verzi programového vybavení, která bude v dalším období průběžně testována a případně upravována v závislosti na problémových oblastech, které budou v rámci řešení projektu zpracovávány.

Publikace:

Fiala D., Jezek, K., Rousellot, F.: Finding Autoritative Researchers on Academic Web Sites, Enformatica, Vol. 17, Dec. 2006, pp. 74 – 79, ISSN 1305 - 5313

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Programové řešení je k dispozici na pracovišti jako autonomní programový subsystém, bohužel k němu zatím (z časových důvodů) nebyla zpracována příslušná dokumentace. Případné získání a využití programového řešení je možné domluvit emailem na adresu dalfia@kiv.zcu.cz. Zdrojové texty programů jsou k dispozici na adrese http://home.zcu.cz/~dalfia/CotSewing/

Číslo aktivity

17

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření nástroje pro zpracování dat dostupných z projektu DMOZ

Zahájení aktivity

3.7.2006

Ukončení aktivity

15.12.2006

Popis aktivity

Projekt DMOZ (detaily viz http://dmoz.org) v současné době obsahuje roztříděné odkazy na webové stránky s ohledem na jejich tematické zaměření. Text, který lze z těchto stránek získat, je velmi vhodný k natrénování klasifikátorů a rozpoznávacích nástrojů, které umožní již automatické vyhledání nových, tématický podobných stránek, případně je vhodný k vytváření univerzálně využitelných námětových korpusů. K dispozici jsou v tomto ohledu i jiné zdroje, jako například adresářové služby serverů Google nebo Yahoo, označované jako Google Directory, respektive Yahoo Directory. Po jejich analýze bylo rozhodnuto využít raději data z projektu DMOZ, jelikož všechny odkazy zde jsou kategorizovány výhradně lidmi, což u adresářových služeb nemusí být pravidlem. V projektu DMOZ proto očekáváme lepší kvalitu dat. V rámci této aktivity tedy byla vytvořena aplikace schopná připojit se k serveru dmoz.org a uložit veškeré zde dostupné datové sety na lokální počítač. Z nich je možné si následně zvolit kategorie témat, ze kterých má být vytvořen námětový textový korpus. Výhody představuje i

možnost specifikovat jazyk, v němž má být požadovaný korpus vytvořen a skutečnost, ze lze specifikovat libovolné tématické úrovně, které mají být brány v úvahu.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná poskytnout požadované výstupy – tématické datové korpusy v požadované kvalitě, v zadaném jazyce.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola vygenerovaných datových korpusů, jejich použití ke klasifikaci – obdržení očekávaných výsledků.

Číslo aktivity

18

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření rozhraní ke klasifikátoru SVM

Zahájení aktivity

4.9.2006

Ukončení aktivity

Popis aktivity

Klasifikátor SVM je v současné době jediný klasifikátor, který dosahuje vynikajících výsledků prakticky ve všech oblastech zpracování dat, včetně zpracování textu. Stalo se proto jednou z priorit tohoto projektu využít jeho vlastností a prozkoumat jeho efektivitu – například při použití různých modelů charakterizujících textové dokumenty. Protože je již na internetu k dispozici jeho univerzální implementace (viz http://svmlight.joachims.org), bylo v rámci této aktivity vytvořeno aplikační rozhraní, které umožňuje použití různých technik využívaných při zpracování textu současně s klasifikátorem SVM. Účelem rozhraní je zjednodušit použití klasifikátoru SVM pro potřeby zpracování textu a odstranit některé kroky, které musejí být díky snaze o co jeho nejuniverzálnější použití v různých oblastech prová-děny. K dispozici je díky vytvořenému rozhraní například automatická úprava umožňující klasifikovat dokumenty do více klasifikačních tříd současně, k-cross fold validace, vyhodnocení celkové úspěšnosti ve formě micro-F1, macro-F1, BEP a dalších měr atd.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity

19

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření webového spidera

Zahájení aktivity

3.7.2006

Ukončení aktivity

10.12.2006

Popis aktivity

Součástí připravovaného systému pro procházení Webu a automatickou identifikaci internetových stránek podle

tématu, o kterém pojednávají, bylo v první fázi vytvoření aplikace označované jako webový spider. Jejím cílem je zadanou výchozí stránku zpracovat, získat z ní nebo určit předem definované údaje včetně odkazů na další stránky, které budou následně stejným způsobem zpracovány. Důležitou vlastností této aplikace je především její modularita, která jednoduchým způsobem dovoluje snadnou modifikaci. Možné využití tedy nepředstavuje jen automatické vytváření námětových korpusů, ale ve spojení s dalšími vhodnými moduly představuje vhodný prostředek pro ověření navržených algoritmů určených přímo k práci s webovým prostředím. Ve spojení s vhodným analyzátorem dat získaných z webového spidera je možné kompletně mapovat určitou tématickou doménu, v našem případě například servery obsahující závadné (protizákonné) materiály, které se vykytují v rámci určitého území – například České republiky nebo států Evropské unie. Neocenitelnými údaji budou bezpochyby i statistiky počtu výskytů určitých slov na závadných webových stránkách, údaje o často se zde vyskytujících emailových adresách, analýza významnosti jednotlivých webových serverů z pohledu množství závadných dat, a podobně.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - modulárně členěná, možnost různých nastavení.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, nastavení konfiguračních parametrů a testování běhu aplikace, zakomponování nově vytvořeného modulu, kontrolní úprava funkčnosti stávajících modulů

Číslo aktivity

20

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Implementace klasifikační metody Naive Bayes rozšířené o současné zpracování bigramů a 2-itemsetů

Zahájení aktivity

1.8.2006

Ukončení aktivity

Popis aktivity

Vlastnosti metody Naive Bayes a její úspěšnost při klasifikaci textových dokumentů byly poměrně podrobně prozkoumány. Stále se zde však objevují nové prostory pro zkoušení nových přístupů vedoucích ke zlepšení dosahovaných výsledků. Jedním z nich je i možnost obohatit obecně používaný bag-of-words model dokumentu (tedy model založený na jen jednotlivých slovech) o další položky. Těmi mohou být například slovní n-gramy nebo itemsety. Doposud byly n-gramy a itemsety zkoumány vždy odděleně, naším cílem se však stalo nejen ověřit, které z obou přístupů poskytují lepší výsledky, ale navíc se i pokusit oba přístupy k zlepšení úspěšnosti klasifikace zkombinovat za účelem vylepšení dosavadních výsledků. A právě uskutečněním této aktivity se nám otvírá možnost tyto experimenty realizovat. Částečně zde byly využity poznatky a principy použité ve výše popsané aktivitě b), zejména se jednalo o fázi vyhodnocení celkové úspěšnosti klasifikace. Díky dokončení této aktivity jsme již měli možnost dosáhnout poměrně zajímavých výsledků, které jsme úspěšně publikovali.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy, možnost různých nastavení nutných pro experimenty.

Publikace:

Tesar R., Poesio M., Strnad V., Jezek K.: “Extending the Single Words-Based Document Model: A Comparison of Bigrams and 2-Itemsets”. The 2006 ACM Symposium on Document Engineering (DocEng’06), Amsterdam, Netherlands, ACM press, ISBN 1-59593-515-0, pages 138-146, http://doi.acm.org/10.1145/ 1166160.1166197, October 2006.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity
21
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Vytvoření vícejazyčného závadného datasetu (pornografie)
Zahájení aktivity
3.7.2006
Ukončení aktivity
15.12.2006
Popis aktivity
Textové klasifikátory vyžadují ke své činnosti obecně dokumenty patřící do rozpoznávané množiny i dokumenty z množiny opačné. Jen tak lze zajistit jejich optimální fungování. Proto byl za pomoci dobrovolníků a nástrojů vytvořených v rámci aktivit 2006-17 a 2006-19 vytvořen dataset z textových dokumentů pokrývající oblast pornografie, a to v českém, slovenském, francouzském a německém jazyce. V souladu s potřebami klasifikátorů textu je ke každému jazyku vždy k dispozici jak množina závadných, tak i nezávadných dokumentů.
Skutečné Indikátory dosažení - výsledky aktivity
Vícejazyčný dataset pokrývající závadnou oblast pornografie a erotiky, obsahuje stejný poměr závadných i nezávadných dokumentů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Vizuální kontrola datasetu, otestování praktické při klasifikaci textu, správná kategorizace je prakticky zajištěna ručním vytvářením korpusu.

Číslo aktivity
22
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Vytvoření korpusu vhodného pro testování prototypových řešení
Zahájení aktivity
3.7.2006
Ukončení aktivity
1.11.2006
Popis aktivity
Pro ověření kvality navrhovaných algoritmů, především z cíle 2, je nutné vytvořit datové korpusy. Datové korpusy jsou v tuto chvíli dvoujazyčné, do budoucna uvažujeme o postupném rozšiřování o další evropské jazyky. Skládají se z vybraných článku tiskových agentur Reuters (anglická část) a ČTK (česká část). Z obou kolekcí byly vybrány příspěvky shodně tématicky zaměřené, což je podmiňující pro některé zkoumané aplikační oblasti. Výsledná kolekce obsahuje 5 tématických tříd po 8000 dokumentech v každé jazykové části. Kolekce je využitelná pro trénování klasifikátorů, testování disambiguace, sumarizaci, případně jako testovací data pro vyhledávací úlohu.
Skutečné Indikátory dosažení - výsledky aktivity
Byl vytvořen korpus českých a anglických tiskových zpráv, obsahující celkem 16000 dokumentů, dělený do 5 tříd

tematicky shodných pro každý jazyk.

Publikace:
Toman M., Tesar R., Jezek K.: “Influence of Word Normalization on Text Classification”. The 1st International Conference on Multidisciplinary Information Sciences & Technologies (InSciT 2006), Merida, Spain, ISBN 84-611-3105-3, pp. 354-358, October 2006.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Korpus byl použit v několika multilingválních úlohách, viz např. 24, 25.

Číslo aktivity
23
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Návrh systému pro tvorbu korpusů z webu
Zahájení aktivity
15.8.2006
Ukončení aktivity
30.11.2006
Popis aktivity
Web je jedním z hlavních zdrojů informace. Pro získávání datových korpusů z prostředí webu byl vytvořen aplikační nástroj založený na použití pravidel. Jednotlivá pravidla definují požadované datové soubory, které jsou následně použity k vytvoření textového korpusu. Vytvořené datové korpusy mohou být libovolně tématicky zaměřené a strukturované v závislosti na použitém zdroji dat. Výhodou řešení je jeho obecnost a robustnost, umožňuje dávkové získání libovolného množství dokumentů z webového prostředí. Tyto dokumenty jsou nutné pro ověření navrhovaných metodik. Aplikace bude využita při tvoření textových korpusů pro další evropské jazyky, jak bylo zmíněno v rámci popisu aktivity 22.
Skutečné Indikátory dosažení - výsledky aktivity
Bylo navrženo a implementováno modulární řešení pro získávání datových korpusů z webu.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Pomocí systému byl vytvořen zkušební korpus z webového serveru německé tiskové agentury Die Welt.

Číslo aktivity
24
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Vytvoření prototypového systému pro vyhledávání
Zahájení aktivity
17.7.2006
Ukončení aktivity

Popis aktivity
V rámci projektu jsme navrhli prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jádrem vyhledávání je thesaurus EuroWordNet a sumarizátor je založen na latentní sémantické analýze. V programovém řešení jsme se zaměřili na zpracování anglického a českého jazyka, nicméně princip zpracování zůstává stejný i pro ostatní jazyky poskytované tezaurem EWN. Součástí systému je modul sloužící pro rozšiřování uživatelských dotazů.

Skutečné Indikátory dosažení - výsledky aktivity

Navržený systém byl otestován na korpusu vytvořeném v rámci plnění cíle 22. Prvotní výsledky byly publikovány na mezinárodní konferenci ELPUB 2006. Touto oblastí se budeme zabývat i v následujících letech a navážeme na výsledky získané z prototypového řešení.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systém byl testován na relevanci získávaných výsledků jak pro české a anglické prostředí, tak i při křížovém zpracování. V takovém případě mohou být výsledné dokumenty napsány v jiném jazyce než byl jazyk dotazu. Výsledky jsou slibné s přesností vyhledávání přes 90 %. Bylo provedeno také srovnání výsledků s přístupem aplikovaným ve vyhledávači Google.

Číslo aktivity

25

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Návrh metod disambiguace v multijazykovém prostředí

Zahájení aktivity

15.9.2006

Ukončení aktivity

Popis aktivity

Rozlišení významů slova je nezbytným krokem pro většinu aplikací zpracovávajících přirozený jazyk. Jedná se o klíčovou úlohu pro správné porozumění sdělení, uplatňuje se v komunikaci člověk-počítač. V této fázi projektu chápeme disambiguaci jako klasifikaci víceznačného slova do tříd, které představují vždy jeden význam slova. Pro řešení problému jsme zvolili bayesovský klasifikátor a modifikovali jeho činnost abychom dosáhli co možná nejlepších výsledků.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikační modul schopný provádět disambiguaci slov v anglickém a českém jazyce. Přesnost výsledků silně závisí na kvalitě trénovacích dat disambiguátoru.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky systému byly testovány na vybraných větách a byly porovnávány výsledky systému s očekávaným výsledkem.

Číslo aktivity

26

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh klasifikačních metod v multijazykovém prostředí

Zahájení aktivity

10.10.2006

Ukončení aktivity

Popis aktivity

Vzhledem k rozšiřující se mezinárodní kooperaci nabývá na důležitosti zpracování dokumentů v různých jazycích.

Proto jsme se rozhodli klasifikační metody rozšířit a modifikovat pro použití v multilingválním prostředí. Cílem experimentů bylo ověřit vliv multilinguality textových korpusů na kvalitu klasifikace a navrhnout vhodné využití tezauru za účelem zlepšení, či zobecnění možností klasifikace multilingválních textových korpusů.

Skutečné Indikátory dosažení - výsledky aktivity

Při použití vhodných klasifikačních algoritmů lze provádět klasifikaci dokumentů zcela nezávislou na jazyku, systém byl ověřen na korpusu získaném v rámci aktivity 2006-22.

Publikace:

Toman M., Tesar R., Jezek K.: “Influence of Word Normalization on Text Classification”. The 1st International Conference on Multidisciplinary Information Sciences & Technologies (InSciT 2006), Merida, Spain, ISBN 84-611-3105-3, pages 354-358, October 2006.

Toman, M.; Steinberger J.; Jezek K.: Searching and Summarizing in a Multilingual Environment, ELPUB2006 – Proceedings of the 10th International Conference on Electronic Publishing, Bansko, Bulgaria, 2006, ISBN 978-954-16-0040-5, 2006, pp. 257-266.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systém byl dosud testován ve dvoujazyčném prostředí čeština – angličtina, kde poskytoval slibné výsledky. Metody jsou navrženy s ohledem na možnost jednoduchého rozšíření
především na významné evropské jazyky.

Číslo aktivity

27

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava sumarizačních kolekcí – single-dokument sumarizace

Zahájení aktivity

3.7.2006

Ukončení aktivity

31.10.2006

Popis aktivity

Pro testování kvality navržených sumarizačních metod bylo nejprve třeba upravit formát standardních sumarizačních korpusů, resp. vytvořit nové. Byl vytvořen XML single-dokument sumarizační formát. Existující kolekce CAST a DUC2002, které obsahují páry dokument-abstrakt v angličtině, byly do tohoto formátu převedeny. Pro testování v češtině byla vytvořena nová kolekce, která obsahuje dokumenty (novinové články) ze zdroje idnes.cz. Pro každý dokument jsou k dispozici extrakty o třech různých délkách – 10%, 20% a 30% slov plného textu. 131 dokumentů bylo anotováno pěti anotátory.

Skutečné Indikátory dosažení - výsledky aktivity

Byl vytvořen XML single-dokument sumarizační formát. Kolekce CAST a DUC-2002 byly převedeny do tohoto formátu. Byl vytvořena a anotována kolekce 131 dokumentů v češtině.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Korpusy byly použity při testování sumarizátoru založeného na LSA.

Číslo aktivity

28

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava sumarizačních kolekcí – multi-dokument sumarizace

Zahájení aktivity

1.11.2006

Ukončení aktivity

Popis aktivity

Pro vývoj multi-dokumentového sumarizátoru je nutné mít k dispozici korpus, který obsahuje shluky dokumentů. Dokumenty shluku se vždy týkají určité události. Tento korpus by měl navíc obsahovat abstrakty shluků o různých délkách. Nejprve byl vytvořen XML formát pro ukládání vhodné struktury pro multi-dokument sumarizaci. Standardně používaná kolekce DUC-2002, která obsahuje 60 shluků dokumentů, byla převedena do tohoto formátu.

Skutečné Indikátory dosažení - výsledky aktivity

Byl vytvořen XML multi-dokument sumarizační formát. Korpus DUC-2002 byl převeden do tohoto formátu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Korpusy budou použity při testování sumarizátoru založeného na LSA.

Číslo aktivity

29

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Experimentální verze sumarizátoru založeného na LSA

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

LSA (latentní sémantická analýza) je proces, který umožňuje nalézt hlavní témata analyzovaného textu na základě společného výskytu termů ve větách textu. Tato vlastnost je využita v navrženém sumarizátoru textů, který extrahuje věty s vysokým obsahem hlavních témat sumarizovaného textu. Ve spolupráci s univerzitou v Essexu (Anglie) byl sumarizátor rozšířen o znalost anaforických řetězců. Pro rezoluci anafor byl využit systém GuiTAR, který byl vyvinut v Anglii. Hlavní témata jsou tedy tvořena nejen z lexikální stránky textu, ale navíc z entitní stránky (entity, které se opakují pomocí anaforických odkazů). Dále byla zkoumána možnost detekce nepodstatných vedlejších vět extraktu, která by umožnila další kompresi textu.

Skutečné Indikátory dosažení - výsledky aktivity

Sumarizační systém založený na LSA, rezoluci anafor a kompresi souvětí.

Publikace:

Steinberger, J., Ježek, K.: Sentence Compression for the LSA-based Summarizer. Proceedings of the 7th International Conference on Information Systems Implementation and Modelling, pp. 141-148, MARQ Ostrava, Přerov, Czech Republic, 2006, ISBN 80-86840-19-0.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systém byl testován na anglických textech – korpusy CAST a DUC-2002. V porovnání se sumarizátory, které se zúčastnily DUC-2002 testování, se LSA sumarizátor umístil na druhém místě z celkových 18 soupeřících sumarizátorů. Testování na kolekci českých testů ověřilo schopnost jednoduché adaptace na jiný jazyk (zde nebyla použita resoluce anafor).

Číslo aktivity

30

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Metoda hodnocení kvality sumarizátorů na základě LSA

Zahájení aktivity

1.8.2006

Ukončení aktivity

Popis aktivity

Získání hlavních témat textu díky latentní sémantické analýze lze využít také při ohodnocování kvality extraktů. Myšlenkou metody je, že hlavní téma extraktu by mělo být co nejvíce podobné hlavnímu tématu plného textu. Dále byla zkoumána možnost použití více hlavních témat.

Skutečné Indikátory dosažení - výsledky aktivity

Systém pro hodnocení kvality extraktů založený na LSA.

Publikace:

Toman, M.; Steinberger J.; Jezek K.: Searching and Summarizing in a Multilingual Environment, ELPUB2006 – Proceedings of the 10th International Conference on Electronic Publishing, Bansko, Bulgaria, 2006, ISBN 978-954-16-0040-5, 2006, pp. 257-266.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byla zjišťována korelace mezi manuálně získaným pořadím sumarizátorů (dle anotátorů) a pořadím získaným navrženou automatickou metodou. Výsledky ukázaly vysokou korelaci – 0.86 (korpus DUC-2002).

Číslo aktivity

31

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Aplikace pro anotaci textů a hodnocení sumarizačních metod

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

Existuje velké množství metod pro hodnocení kvality extraktů. Každá skupina metod má své přednosti a také slabiny. Vytvořená aplikace umožňuje jednoduše anotovat texty pro sumarizaci (anotátor vybírá nejvýznamnější věty textu). Výběr „ideálních“ vět je však činnost velice subjektivní. Systém umožňuje měřit shodu anotátorů – metody Percent Agreement a Kappa. Další funkcí systému je připojení kolekce extraktů a hodnocení jejich kvality následujícími metodami: Precision, Recall, F-score, Relative Utility a Cosine Similarity.

Skutečné Indikátory dosažení - výsledky aktivity

Aplikace pro anotaci textů a hodnocení sumarizačních metod.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

V této aplikaci byla anotována kolekce českých textů (viz aktivita 2006-27). Průměrná shoda anotátorů byla 0.4 (Kappa). Implementované metody hodnocení extraktů byly také použity pro srovnání výsledků s metodou založenou na LSA.

Číslo aktivity

32

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vývoj nových metod filtrace a klasifikace textů a webových stránek, včetně testování prototypů klasifikátorů a jejich aplikace v aktuálních oblastech jako jsou např. spamové analýzy nebo zjišťování webových stránek s protizákonnou tématikou.

Zahájení aktivity

1.7.2006

Ukončení aktivity

31.12.2006

Popis aktivity

Pro potřeby vývoje metod klasifikace textů byla shromážděna rozsáhlá sada dokumentů, a to již částečně klasifikovaných dle normy MSC 2000 (Mathematical Subject Classification) a provedeno její předzpracování pro potřeby strojového učení. Jde o články digitalizované v rámci projektu DML CZ [1].

Skutečné Indikátory dosažení - výsledky aktivity

[1] Bartošek, Miroslav - Lhoták, Martin - Rákosník, Jiří - Sojka, Petr - Šárfy, Martin. DML-CZ: The Objectives and the First Steps. A.K. Peters Ltd., 14 pp., accepted for publication, 2007.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

---

Číslo aktivity

33

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vývoj nových metod filtrace a klasifikace textů a webových stránek, včetně testování prototypů klasifikátorů a jejich aplikace v aktuálních oblastech jako jsou např. spamové analýzy nebo zjišťování webových stránek s protizákonnou tématikou.

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

Byla věnována pozornost budování první části systému pro detekci plagiátů – hrubé vyhledání množiny podobných dokumentů z potenciálně velkého množství (škálovatelnost). Řešené problémy: - segmentace textu – rozsekání dokumentů na tematicky ucelené části. Implementace nové metody pro segmentaci textu, založené na LSI transformaci.Srovnání s Hearstovou metodou TextTiling. - indexování – nalezení tematicky podobných segmentů v potenciálně velkém množství. - - - srovnání naivního porovnávání, kd-tree a transformace pomocí spektrální analýzy spolu se shlukováním k-means. - fulltextové srovnání dvou textů - nalezení a provázání (pomocí XML značek) shodných úseků mezi dvěma texty. - PANK (algoritmus pro výběr podmnožiny rysů založený na

beam-search; dokončení a optimalizace implementace, testy na databázi s proteiny. Práce na publikaci shrnující výsledky. - výchozí algoritmus pro detekci plagiátů a ověřován na vzorcích dat z Informačního systému MU (IS). Modifikovaná a rozšířená implementace se aktuálně využívá na plných datech IS MU (zprávy o tom byly publikovány v denním tisku.

Skutečné Indikátory dosažení - výsledky aktivity

Publikace o vlivu parametrů předzpracování na klasifikaci textu, která byla přijata na mezinárodní konferenci ICCSSE12
v Bangkoku konané v lednu 2007.
Pomikálek, J. Řehůřek, R., The Influence of Preprocessing
Parameters on Text Categorization, paper accepted at XIX Int.
Conference on Computer and Systems Science
and Egineering, January 29-31, Bangkok 2007.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Implementovaný algoritmus.

Číslo aktivity

34

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vývoj metod, návrh a implementace prototypu kategorizačního a vyhledávacího systému pro multilinguální prostředí, včetně prostředí českého jazyka a jeho propojení s dalšími jazyky.

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

Byl systematicky doplňován český WordNet o tzv. hloubkové valenční rámce ve vazbě na základní princetonskou verzi anglického WordNetu (PWN 2.0). V rámci doplňování probíhala také kontrola zejména slovesných synsetů a jejich opravy. Vzhledem k tomu, že PWN 2.0 je prostřednictvím tzv. mezijazykového indexu (ILI) napojen na nejméně 13 evropských jazyků, jsou tyto vazby naprosto nezbytné pro vývoj předpokládaného multilinguálního prostředí.

Skutečné Indikátory dosažení - výsledky aktivity

Aktuální verze českého WordNetu přístupná pod editorem DebVisdic.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

---

Číslo aktivity

35

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Práce na lexikální databázi valenčních rámců českých sloves Verbalex

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

Pracovali jsme na tvorbě lexikální databáze Verbalex obsahující valenční rámce českých sloves, které jsou formálním prostředkem pro zachycení sématické struktury promluv a také její rozpoznávání. Databáze Verbalex čítá nyní cca 11 000 českých sloves. Byla též řešena problematika inventáře tzv. ontologických kategorií, které se objevují v rámcích na místě argumentů slovesných predikátů. Výsledkem je inventář sémantických rolí, který čítá celkem cca 250 ontologických kategorií, jež zde slouží jako sémantické role umožňující rozpoznávat sémantiku vět vyskytujících se v běžných textech.

Skutečné Indikátory dosažení - výsledky aktivity

Databáze Verbalex je k dispozici pod námi vytvořeným prohlížečem GVIM.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výchozí verze databáze Verbalex je přístupná na adrese http://nlp.fi.muni.cz/verbalex/html/generated/alphabet/index-A.html

---

Číslo aktivity

36

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vývoj a analýza dialogových rozhraní založených na sémantickém webu pro aplikace v asistivních technologiích, se zvláštním zřetelem na informační dostupnost (accessibility) a možnosti vytvářet webové stránky pomocí dialogu pro nevidomé.

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

V tomto bodě jsme se zaměřili na analýzu problému vytváření webovských stránek a počítačové grafiky v kontextu sémantického webu a s aplikacemi v asistivních technologiích, zejména s ohledem na zpřístupňování informací a možnosti vytváření internetových prezentací pro nevidomé. Byl vypracován a publikován základní koncept takového přístupu, který současně splňuje a zajišťuje požadavky přístupnosti vůči nevidomým (internetový standard Web Content Accessibility). Současně byl vypracován a publikován koncept generování grafických objektů, založený na formální bázi Pawlakových informačních systémů popisujících elementární ontologie, který do formátu SVG integruje popis odpovídajícího grafického objektu, odvozený z historie jeho dialogového vygenerování. Tento přístup umožňuje snadno vytvářet grafiku, jejíž popis je čitelně zahrnut do formátu grafického objektu takovým způsobem, že jej lze využívat k automatickému získání popisu vygenerované grafiky na webovských presentacích, které se tak stávají přístupnější pro nevidomé uživatele. V souvislosti s tímto aspektem byla navržena a částečně implementována metoda pro efektivní získávání informací o vygenerovaných grafických objektech.

Skutečné Indikátory dosažení - výsledky aktivity

Publikace:

Kopeček, Ivan - Ošlejšek, Radek. Creating Pictures by Dialogue. In Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006. Berlin : Springer-Verlag, 2006,. s. 61-68, 8 s. ISBN 3-540-36020-4

Kopeček, Ivan - Ošlejšek, Radek. The Blind and Creating Computer Graphics. In Proceedings of the Second IASTED International Conference on Computational Intelligence. Anaheim, Calgary, Zurich: ACTA Press, 2006,. s. 343-348, 6 s. ISBN 0-88986-602-3

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Viz příslušné sborníky.

Číslo aktivity

37

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vývoj nových metod pro reprezentaci znalostí na základě temporálních intenzionálních logik vyšších řádů.

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

Transparentní intenzionální logika (TIL) poskytuje mocný aparát pro zachycení významu vět přirozeného jazyka. Proti jiným logikám TIL umožňuje korektně zpracovat časový aspekt, intenzionalitu a další obtížně zachytitelné rysy typické pro objekty popisované přirozeným jazykem. V rámci projektu v současnosti vyvíjíme techniky pro efektivní zpracování báze znalostí (ukládání, propojení, vyhledávání relací a kontrola konzistence vložených faktů) založené na TIL. Součástí těchto technik jei základní inferenční stroj umožňující nalézat odpovědi s využitím jednoduchého vyvozování. Návrh báze znalostí, na rozdíl od dřívějšího pokusu o implementaci podobného systému (Chrz, 1984), od začátku počítá se zpracováním času v analyzovaných větách. Vlastní vkládaná báze znalostí má podobu datového zdroje nezávislého na konkrétním přirozeném jazyce.

Skutečné Indikátory dosažení - výsledky aktivity

Rozpracovaný systém Dolphin

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Poster pro konferenci ...

Číslo aktivity

38

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Budování modulu pro rozpoznávání anaforických vztahů v češtině.

Zahájení aktivity

Ukončení aktivity

Popis aktivity

Věnovali jsme zejména tvorbě nástrojů na zpracování dat z Pražského závislostního korpusu (PDT – Prague Dependency Treebank). Pracovalo se nástrojích pro načítání těchto dat ve fs-formátu, dále se testovaly moduly pro detekci referujících výrazů a anafor. Zejména detekce anafor je poměrně komplexní proces, pro který zdaleka ne všechny potřebné údaje jsou obsaženy v anotaci PDT. Například ke korektní detekci nevyjádřených subjektů je klíčové správné rozčlenění závislostních stromů do klauzí, nalezení predikátu a určení jeho morfologických kategorií. Dalším obtížným krokem bylo zarovnávání (alignment) koreferenční anotace obsažené v PDT - k tomu slouží detektory tak, aby bylo možné vyhodnocovat výsledky algoritmů implementovaných v systému. Pracovalo se též na metrikách pro vyhodnocování úspěšnosti AR algoritmů,konkrétně se vyhodnocovaly se re-implementované algoritmy (Hajičová 1987, Hobbs' syntactic search, Centering algorithm) a analýzou chyb se dospívá k úpravám zlepšujícím jejich úspěšnost. Dále jsme věnovali pozornost nástrojům pro práci s dalšími formáty dat (zatím z korpusů PDT a TigerXML (němčina)), rozpracovány jsou úpravy analyzátorů Synt a Dis a re-implemenace algoritmu inspirovaného systémem Lappina a Leasse. V tomto roce se pozornost věnovala zejména tvorbě nástrojů na zpracování dat z Pražského závislostního korpusu (PDT – Prague Dependency Treebank). Pracovalo se nástrojích pro načítání těchto dat ve fs-formátu, dále se testovaly moduly pro detekci referujících výrazů a

anafor. Zejména detekce anafor je poměrně komplexní proces, pro který zdaleka ne všechny potřebné údaje jsou obsaženy v anotaci PDT. Například ke korektní detekci nevyjádřených subjektů je klíčové správné rozčlenění závislostních stromů do klauzí, nalezení predikátu a určení jeho morfologických kategorií. Dalším obtížným krokem bylo zarovnávání (alignment) koreferenční anotace obsažené v PDT - k tomu slouží detektory tak, aby bylo možné vyhodnocovat výsledky algoritmů implementovaných v systému. Pracovalo se též na metrikách pro vyhodnocování úspěšnosti AR algoritmů, konkrétně se vyhodnocovaly se re-implementované algoritmy (Hajičová 1987, Hobbs' syntactic search, Centering algorithm) a analýzou chyb se dospívá k úpravám zlepšujícím jejich úspěšnost. Dále jsme věnovali pozornost nástrojům pro práci s dalšími formáty dat (zatím z korpusů PDT a TigerXML (němčina)), rozpracovány jsou úpravy analyzátorů Synt a Dis a re-implemenace algoritmu inspirovaného systémem Lappina a Leasse a také další algoritmus vyvíjený ve spolupráci s ÚFAL MFF UK. V tomto roce se pozornost věnovala zejména tvorbě nástrojů na zpracování dat z Pražského závislostního korpusu (PDT – Prague Dependency Treebank). Pracovalo se nástrojích pro načítání těchto dat ve fs-formátu, dále se testovaly moduly pro detekci referujících výrazů a anafor. Zejména detekce anafor je poměrně komplexní proces, pro který zdaleka ne všechny potřebné údaje jsou obsaženy v anotaci PDT. Například ke korektní detekci nevyjádřených subjektů je klíčové správné rozčlenění závislostních stromů do klauzí, nalezení predikátu a určení jeho morfologických kategorií. Dalším obtížným krokem bylo zarovnávání (alignment) koreferenční anotace obsažené v PDT - k tomu slouží detektory tak, aby bylo možné vyhodnocovat výsledky algoritmů implementovaných v systému. Pracovalo se též na metrikách pro vyhodnocování úspěšnosti AR algoritmů, konkrétně se vyhodnocovaly se re-implementované algoritmy (Hajičová 1987, Hobbs' syntactic search, Centering algorithm) a analýzou chyb se dospívá k úpravám zlepšujícím jejich úspěšnost. Dále jsme věnovali pozornost nástrojům pro práci s dalšími formáty dat (zatím z korpusů PDT a TigerXML (němčina)), rozpracovány jsou úpravy analyzátorů Synt a Dis a re-implemenace algoritmu inspirovaného systémem Lappina a Leasse a také další algoritmus vyvíjený ve spolupráci s ÚFAL MFF UK. Na základě porovnání všech implementovaných algoritmů (částečně s využitím metod strojového učení) je připraven meta-algoritmus, který kombinuje výhody jednotlivých algoritmů a dosahuje tak lepších výsledků.

Skutečné Indikátory dosažení - výsledky aktivity

Připravené publikace

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

---

Číslo aktivity

39

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Vývoj metod citační analýzy, tvorby prototypu umožňujícího vyhledávání citačních komunit a vlivných entit na základě topologie web stránek i referencí z elektronicky publikovaných článků.

Zahájení aktivity

Ukončení aktivity

Popis aktivity

Byl vyvinut prototyp pro detekci seznamu publikací v dokumentech a identifikace jednotlivých položek literatury pomocí regulárních výrazů. Byly testovány algoritmy pro vyhledání citačních záznamů v databázích metadat založené na vážených histogramech četností n-gramů. [1,2,3]

Skutečné Indikátory dosažení - výsledky aktivity

Publikované výsledky:

[1] Bartošek, Miroslav - Lhoták, Martin - Rákosník, Jiří - Sojka, Petr - Šárfy, Martin. DML-CZ: The Objectives and the First Steps. A.K. Peters Ltd., 14 pp., accepted for publication, 2007.

[2] Radovan Panák: Digitalizácia matematických textov. Diplomová práce FI MU, 2006, vedoucí P. Sojka.

[3] Tomáš Mudrák: Digitalizace matematických textů. Diplomová práce FI MU, 2006, vedoucí P. Sojka.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

2.2.2. AKTIVITY NEUSKUTEČNĚNÉ v roce 2006

Číslo aktivity
Ke kterému dílčímu cíli se aktivita vztahuje
Název (cíl)aktivity
Zahájení aktivity
Ukončení aktivity
Popis aktivity
Důvody, proč se aktivitu nepodařilo uskutečnit

2.3.NÁKLADY PROJEKTU - 2006

2.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2006
Typ	skutečné
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1394**	**1379**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**1094**	**640**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**50**	**25**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**22**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**40**	**26**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**98**	**49**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**160**	**0**
F9. CELKEM	**2858**	**2119**

Rok	2006
Typ	skutečné
Organizace	Masarykova univerzita
Role organizace	spolupříjemce

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**912**	**755**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**75**	**75**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**30**	**20**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**30**	**20**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**50**	**27**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**100**	**0**
F9. CELKEM	**1197**	**897**

2.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2006
Typ	skutečné
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**2306**	**2134**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**1169**	**715**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**30**	**20**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**80**	**45**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**22**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**40**	**26**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**148**	**76**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**260**	**0**
F9. CELKEM	**4055**	**3016**

2.3.3. ZDŮVODNĚNÍ ZMĚN V ČERPÁNÍ

Nedočerpáno zůstalo 97 tisíc Kč mzdových prostředků, neboť dva doktorandi, kteří v létě ukončili studium a s nimiž bylo při plánování projektu počítáno jako s vědeckovýzkumnými pracovníky, se na základě změněné rodinné situace (oženili se) rozhodli z univerzity odejít. Jelikož své rozhodnutí oznámili až v červnu 2006, nepodařilo se okamžitě za ně

získat náhradu. Pro řešení plánovaných prací byli za ně přijati jiní pracovníci (Š. Albrecht, P.Král, J.Hynek), avšak s jistou časovou prodlevou nutnou pro výběrové a administrativní úkony. Jeden z členů kolektivu (D. Fiala), který je v doktorandském studiu pod dvojím vedením, byl v říjnu povolán francouzským spoluškolitelem na univerzitu ve Štrasburku. I když ve Francii pracuje na úkolech souvisejících s výzkumem našeho projektu, neměli jsme dostatek prostředků na jeho vyslání formou dlouhodobé zahraniční pracovní cesty. Požádal proto na doporučení právního odboru univerzity o neplacené pracovní volno a zbylé finanční prostředky jsou zahrnuty v převáděné částce.

Převedená částka bude v roce 2007 použita k úhradě zvýšených osobních nákladů na potřebné intenzivnější zapojení studentů magisterského studia formou dohod o provedení práce, pro nově přijaté a ze zahraničních stáží se vracející pracovníky.

Veškeré ostatní finanční prostředky byly vyčerpány v souladu s plánem.

2.3.4. NEVYUŽITÉ FINANČNÍ PROSTŘEDKY

Nevyužité mzdové prostředky v částce 97 tisíc Kč byly převedeny do fondu účelově určených prostředků s tím, že budou využity v roce 2007 na podporu činnosti nově přijatých doktorandů, kteří jsou na řešení projektu již připravováni. Převedené prostředky umožní pracovišti přijmout o jednoho doktoranda více, což bude mít dva pozitivní efekty - napomůže mladým pracovníkům vytvořit při řešení projektu žádoucí konkurenční prostředí a poskytne mladým talentovaným absolventům možnost v širší míře se uplatnit při řešení progresivních vědeckovýzkumných úloh.

2.3.5. Seznam hmotného a nehmotného majetku pořízeného za sledované období

Pořadí	1
Název	Server DELL PE6850 (4x Dual Core Xeon 7140M, 3,4GHz, 16MB cache, 3x 73GB SAS 15000rpm, 32GB DDR2 RAM, RAID-5 controller, DVD-ROM, DRAC 4 SMC, APC UPS 3000i 2U)
Podíl užití majetku pro řešení v %	100
Pořizovací cena v tis. Kč	1093
Uznaný náklad v tis. Kč	1093
Uhrazeno z dotace v tis. Kč	640
Datum dodání	1.12.2006
Datum zprovoznění	14.12.2006
Dodavatel	AXES Computers, s.r.o., Plzeň

Pořadí	2
Název	Paměťový modul DIMM 1024MB DDR2 PC800 Cosair XMS2-6400, Disk Seagate 500GB ST3500630AS 7200.10 SATA 16MB
Podíl užití majetku pro řešení v %	100
Pořizovací cena v tis. Kč	11
Uznaný náklad v tis. Kč	11
Uhrazeno z dotace v tis. Kč	11
Datum dodání	20.10.2006
Datum zprovoznění	22.10.2006
Dodavatel	AXES Computers, s.r.o., Plzeň

Pořadí	3
Název	Experimentální stanice
Podíl užití majetku pro řešení v %	100
Pořizovací cena v tis. Kč	38,5
Uznaný náklad v tis. Kč	38,5
Uhrazeno z dotace v tis. Kč	25
Datum dodání	20.10.2006
Datum zprovoznění	24.10.2006
Dodavatel	AXES Computers, s.r.o., Plzeň

3. ZÁMĚR A NÁVRHY PRO NÁSLEDUJÍCÍ OBDOBÍ - rok 2007

3.1. AKTIVITY PLÁNOVANÉ NA DALŠÍ OBDOBÍ - rok 2007

Číslo aktivity

Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity

Zahájení aktivity

Ukončení aktivity

Popis aktivity

Plánované indikátory dosažení - očekávané výsledky aktivity

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

---

Číslo aktivity
01
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Pořizování korpusu LAC-SS
Zahájení aktivity
2.12.2006
Ukončení aktivity

Popis aktivity
Pořízení a transkripce nahrávek spontánní řeči do korpusu LAC-SS (LICS Audio Corpus – Spontaneous Speech). Postupovat se bude podle metodologie vytvořené v rámci aktivity 2006-01 "Příprava pořizování korpusu LAC-SS2006". Pořizování nahrávek se předpokládá v průběhu semestrů, kdy studenti použití k nahrávání jsou snadno dosažitelní, souběžně s pořizováním, resp. návazně na něj, bude pak prováděna ortografická transkripce jednotlivých nahrávek.
Plánované indikátory dosažení - očekávané výsledky aktivity
Nahrávky a jejich ortografická transkripce.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Nahrávky budou použity k trénování akustických modelů rozpoznávače. Ověřením bude mj. i výsledná úspěšnost rozpoznávání.

Číslo aktivity
02
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Integrace komponent systému LASER

Zahájení aktivity

1.2.2007

Ukončení aktivity

31.12.2007

Popis aktivity

V současné době jsou jednotlivé komponenty rozpoznávače LASER (viz http://liks.fav.zcu.cz/mediawiki/index.php/LASER) implementovány jako programy spustitelné z příkazové řádky, jejichž vstupem a výstupem je datový soubor. Tento způsob implementace je vhodný pro testování a výzkum, nikoli však pro real time aplikaci. Využití souborů jako prostředku pro komunikaci mezi moduly značně prodlužuje dobu odezvy systému, protože 1. čtení souborů z disku je pomalé, 2. ostatní komponenty musejí čekat, než skončí nahrávání (nahrávání není výpočetně náročné a zbylý procesorový čas je možné využít). Řešením proto bude implementace jednotlivých komponent jako procesů se vstupní a výstupní datovou frontou, což představuje poměrně rozsáhlou úpravu experimentálního programového vybavení.

Plánované indikátory dosažení - očekávané výsledky aktivity

Sada knihoven poskytujících funkce jednotlivých komponent rozpoznávače.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Porovnání doby odezvy se souborově orientovaným přístupem, snadnost připojení knihoven do výsledné aplikace.

Číslo aktivity

03

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Detektor ticha a řeči (2)

Zahájení aktivity

2.1.2007

Ukončení aktivity

31.12.2007

Popis aktivity

Navazuje na aktivitu z roku 2006, zahrnuje implementaci rozpoznávacích algoritmů a natrénování modelů. Budou testovány dva odlišné přístupy k problému: využití vhodné umělé neuronové sítě a modelování pomocí směsi Gaussových funkcí.

Plánované indikátory dosažení - očekávané výsledky aktivity

Knihovna programů pro detekci ticha a řeči a sada natrénovaných modelů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Bude vytvořena statistika chyb.

Číslo aktivity

04

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Tvorba anotačních schémat

Zahájení aktivity

1.12.2006

Ukončení aktivity

30.6.2007

Popis aktivity

Anotační schéma přesně určuje všechny možnosti, jak lze větu sémanticky označit. Anotační schéma je

hierarchická stromová struktura, každé anotační schéma definuje anotační značky pro téma, fráze, pod-fráze až k lexikálním třídám. Téma je kořenem stromu anotačního schématu a lexikální třídy jsou jeho listy. Fráze a pod-fráze pak tvoří uzly stromu, které nejsou ani listy, ani kořen. Anotační schéma je naprosto nezbytné pro aktivitu 2007-05 (Sémantické anotování korpusu), jelikož podle anotačního schématu se bude věta anotovat.

Plánované indikátory dosažení - očekávané výsledky aktivity

Anotační schémata pokrývající většinu zodpověditelných dotazů vytvořeného korpusu v aktivitě 2006-04 (Příprava a vytváření korpusu vět z vybraných domén).

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Prostředkem pro ověření je tzv. mezi-anotátorská shoda (inter-annotator agreement), která určuje, do jaké míry se výsledky anotace jednotlivých anotátorů shodují. Víceznačné nebo nepřesné anotační schéma způsobí nízkou mezi-anotátorskou shodu.

Číslo aktivity

05

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Sémantické anotování korpusu

Zahájení aktivity

8.1.2007

Ukončení aktivity

30.9.2007

Popis aktivity

V rámci této aktivity bude korpus získaný v aktivitě 2006-04 (Příprava a vytváření korpusu vět z vybraných domén) sémanticky označkován (určení významu vět). Sémantickou anotaci bude provádět tým odborně vyškolených pracovníků sestavený v aktivitě 2006-07 (Výběr a zaškolení pracovníků provádějící sémantické anotace). Tým anotátorů bude pro svou práci využívat editor anotací získaný aktivitou 2006-02 (Vytvoření software pro editací sémantických anotací). Sémantické značkování bude probíhat podle anotačních schémat navržených v aktivitě 2007-04 (Tvorba anotačních schémat). V první fázi této aktivity dojde k filtraci dotazů a k určení témat. V druhé fázi bude pro tu část vět, která bude použitelná (počítačově zodpověditelná z informací na internetu), určen význam každé věty. Každá věta bude vždy anotována dvěma anotátory. Věty, ve kterých se anotátoři neshodnou bude dále analyzovány. Dokončením této aktivity dojde ke splnění podcíle 1 – e). Korpus bude sice i nadále rozšiřován a následně anotován také po skončení této aktivity. Avšak po jejím skončení by již mělo být k dispozici dostatečné množství kvalitních dat pro vývoj algoritmů automatické sémantické analýzy.

Plánované indikátory dosažení - očekávané výsledky aktivity

Sémantická analýza zodpověditelných vět korpusu určená týmem anotátorů. Indikátorem dosažení bude počet označkovaných vět.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ověřením bude dosažení alespoň 7000 sémanticky označkovaných vět.

Číslo aktivity

06

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj algoritmů pro automatickou identifikaci lexikálních tříd

Zahájení aktivity

1.9.2007

Ukončení aktivity

1.9.2008

Popis aktivity

Pojmem lexikální třída označujeme v této práci skupinu slov nebo slovních spojení, které dohromady spadají pod nějaký nadřazený pojem. Např. lexikální třída MĚSTA obsahuje mimo jiné města Plzeň, Praha, Brno, ... Lexikální třída ČAS zahrnuje fráze 14:30, odpoledne, atd. Lexikální třídy jsou v korpusu anotovány v rámci aktivity 2007-05 (Sémantické anotování korpusu). Cílem této aktivity je vytvořit sadu algoritmických postupů, které umožní automatickou identifikaci lexikálních tříd v textu. Mezi metody, které budou použity se řadí o částečná sémantická analýza bezkontextovou gramatikou, o analýza regulárními výrazy, o slovníkový přístup.

Plánované indikátory dosažení - očekávané výsledky aktivity

Sada algoritmů, které budou schopny automatické identifkace lexikálních tříd v textu. Identfikátorem dosažení je procentuální úspěšnost určení jednotlivých lexikálních tříd.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Dosažení alespoň 70% úspěšnosti určování lexikálních tříd.

Číslo aktivity

07

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Ověření vlastností vhodného typu neuronové sítě pro zpracování sémantiky přirozeného jazyka

Zahájení aktivity

2.1.2007

Ukončení aktivity

22.12.2007

Popis aktivity

Dokončení testů s dostupnými neuronovými simulátory SOMPAK a SOMtoolbox. Nalezení vhodné metriky, kterou využívá Kohonenova mapa pro vytvoření shluků podobných vstupních vektorů. Výpočty ve standardní Kohonenově mapě obvykle využívají euklidovskou metriku k posuzování podobnosti vstupních vektorů. WEBSOM algoritmus vytváří tzv. mapu slovních kategorií, tj. vstupní vektorové reprezentace slov jsou shlukovány na základě podobnosti významu jednotlivých slov. Cílem této aktivity je ověřit, zda je euklidovská metrika použitá ve standardním SOM algoritmu vhodná pro vytvoření mapy slovních kategorií, vzhledem ke zvolenému způsobu kódování slov v dokumentu (viz předchozí bod). Volba velikosti Kohonenovy sítě a vytvoření mapy slovních kategorií (sémantické mapy) natrénováním navržené SOM sítě trénovací množinou sestavenou z připravovaných českých korpusů. V natrénované Kohonenově mapě, odpovídá každá výpočetní jednotka (neuron) množině slov s podobným sémantickým významem. V této části tedy bude důležité zvolit vhodnou velikost mapy, tak aby jednotlivé slovní kategorie odpovídaly slovům s podobným významem z hlediska českého jazyka. Jak již bylo řečeno, v dostupné literatuře jsou publikovány výsledky a postupy aplikované na anglicky psané dokumenty, v případě česky psaných dokumentů bude pravděpodobně nutné provést určitou modifikaci s ohledem na použitý korpus. Porovnání výsledků dosažených použitím Kohonenovy mapy s klasickým přístupem.

Plánované indikátory dosažení - očekávané výsledky aktivity

Vytvoření tzv. mapy slovních kategorií, modifikace slovní mapy s ohledem na použitý korpus.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ověření, zda je euklidovská metrika použitá ve standardním SOM algoritmu vhodná pro vytvoření mapy slovních kategorií vzhledem ke zvolenému způsobu kódování slov v dokumentu, a konečně ověření modifikace slovní mapy s ohledem na použitý korpus.

Číslo aktivity

08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření vhodných vstupů neuronové sítě

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.9.2007

Popis aktivity

Vytvoření vhodného rozhraní mezi dokumenty (textovou informací) a vstupem Kohonenovy mapy. Vzhledem k tomu, že vstup Kohonenovy mapy a výpočty v mapě jsou založeny na numerickém základě, bude nutné vytvořit vhodnou číselnou reprezentaci slov obsažených v dokumentu. Dále bude nutné z této číselné reprezentace zachytit určitým způsobem sémantiku jednotlivých slov. V dostupné prostudované literatuře existují způsoby, jak vhodně kódovat slova a jejich význam numericky. Cílem další aktivity tedy bude ověřit tento způsob kódování na již vytvořených korpusech a vytvořit rozsáhlou trénovací množinu pro Kohonenovu mapu (WEBSOM).

Plánované indikátory dosažení - očekávané výsledky aktivity

Vytvoření vhodného rozhraní mezi dokumenty (textovou informací) a vstupem Kohonenovy mapy.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ověření daného způsobu kódování na již vytvořených korpusech a vytvoření rozsáhlé trénovací množiny pro Kohonenovu mapu (WEBSOM).

Číslo aktivity

09

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Návrh a implementace modelu webgrafu s cílem určit autoritativnost uzlů zohledňující skryté vazby komponent

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.6.2007

Popis aktivity

Data z uznávané digitální knihovny DBLP ve formátu XML převedeme do relační databáze a aplikujeme na ni algoritmy s modifikovaným vzorcem pro výpočet autorit. Tyto metody jsme využívali již v předchozí fázi projektu, ovšem naše data pocházela přímo z webu a nebyla proto zcela přesná a spolehlivá. To nám tedy neumožňovalo srovnávat výsledky s jinými relevantními údaji. Naproti tomu využití zdrojů DBLP takové srovnání velmi usnadňuje.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - možnost různých relevantních nastavení.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola, nastavení konfiguračních parametrů a testování poskytovaných výstupů.

Číslo aktivity

10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Vyhodnocení výsledků navržených algoritmů pro analýzu struktury Webu

Zahájení aktivity

2.5.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Bude zapotřebí provést rozsáhlé srovnávání výsledků našich algoritmů s jinými, již existujícími informacemi. Navíc není úplně jasné jak a s čím srovnávat. Nejprve je tedy nutno vymyslet a navrhnout srovnávací testy. Nabízejí se porovnání autoritativních vědců se složením programových výborů konferencí, edičních rad časopisů, se seznamy držitelů nejrůznějších ocenění v dané oblasti apod.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - možnost různých relevantních nastavení.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Nastavení konfiguračních parametrů a testování poskytovaných výstupů.

Číslo aktivity

11

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

SPOT – nový webový projekt on-line slovníku překladů odborných termínů

Zahájení aktivity

3.1.2007

Ukončení aktivity

20.12.2007

Popis aktivity

Návrh modelu on-line vytváření doménových slovníků, popis datového a funkčního modelu, implementace jednotlivých modulů, vytvoření webového rozhraní (CZ a EN), doplnění inkrementálních funkcí v systému klient server. SPOT bude zaměřen na užší oblast ICT a v jejím kontextu poskytne nástroj pro vytváření kvalitních českých ekvivalentů anglické terminologie. Výsledný slovník odborné terminologie bude sloužit dvěma účelům: jako zdroj referenčních překladů a jako platforma pro diskuse při jejich „ustalování“. SPOT bude rovněž možné nasadit v dalších aplikačních oblastech, jako je sémantický web, multilinguální rozhraní, rozšiřování dotazů v rámci IR, thesaury apod.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity

12

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava sumarizačních kolekcí – multi-dokument sumarizace v češtině

Zahájení aktivity

2.1.2007

Ukončení aktivity

22.12.2007

Popis aktivity

V současné době neexistuje kolekce českých dokumentů anotovaná pro sumarizaci. V rámci tohoto dílčího cíle bude vytvořen korpus, který bude obsahovat shluky dokumentů. Dokumenty shluku se budou vždy týkat určité události. Pro každý dokument a shluk budou anotátory vytvořeny jak abstrakty (mohou obsahovat nové věty) tak extrakty (obsahují pouze věty původního textu). Vše bude ukládáno do již vytvořeného XML formátu.

Plánované indikátory dosažení - očekávané výsledky aktivity

Český multi-dokument sumarizační korpus.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Použití při testování sumarizátoru založeného na LSA a dalších baseline sumarizátorů.

Číslo aktivity

13

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Experimentální verze sumarizátoru založeného na LSA – multi-dokument sumarizace

Zahájení aktivity

2.1.2007

Ukončení aktivity

22.12.2007

Popis aktivity

Latentní sémantická analýza byla úspěšně aplikována na single-dokument sumarizaci. Dalším postupným cílem je aplikování podobných principů na multi-dokument sumarizaci – sumarizovat se bude celý shluk dokumentů. Rezoluce anafor se již bude muset řešit skrz více textů najednou. Dále bude zkoumána možnost detekce nepodstatných vedlejších vět extraktu, která by umožnila další kompresi textu.

Plánované indikátory dosažení - očekávané výsledky aktivity

Experimentální systém multi-dokument sumarizace textů založený na LSA, rezoluci anafor a kompresi souvětí.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testování na DUC korpusech a nově vytvořené kolekci českých textů.

Číslo aktivity

14

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Metoda hodnocení kvality sumarizátorů na základě LSA

Zahájení aktivity

2.1.2007

Ukončení aktivity

22.12.2007

Popis aktivity

Myšlenkou metody je, že hlavní témata extraktu by měly být co nejvíce podobné hlavním tématům plného textu, popřípadě abstraktu. Zde je nutné nalézt vhodné schéma vážení termů. Dále také testování rozšířit o další hodnotící metody, se kterými by bylo možné navrženou metodu porovnat.

Plánované indikátory dosažení - očekávané výsledky aktivity

Systém pro hodnocení kvality extraktů založený na LSA.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Zjištění korelace mezi manuálně získaným pořadím sumarizátorů (dle anotátorů) a pořadím získaným

automatickými metodami.

Číslo aktivity

15

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Aplikace pro anotaci textů a hodnocení sumarizačních metod

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Aplikace bude rozšířena pro práci s multi-dokument sumarizačními korpusy. Navíc bude umožňovat tvorbu abstraktů. Dále zde budou implementovány varianty ROUGE hodnocení.

Plánované indikátory dosažení - očekávané výsledky aktivity

Rozšíření stávající aplikace o nové metody a práci s multi-dokument sumarizačními korpusy.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Použití při anotaci českých textů. Testování také na DUC kolekcích.

Číslo aktivity

16

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření systému pro vyhledávání

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

V rámci projektu navrhujeme prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jádrem vyhledávání je thesaurus EuroWordNet a sumarizátor je založen na latentní sémantické analýze. V současném řešení se zaměřili na zpracování anglického a českého jazyka. Jelikož princip zpracování zůstává stejný i pro ostatní jazyky poskytované tezaurem EWN, plánujeme vyzkoušet postup i pro další evropské jazyky. Hlavní důraz bude kladen na zvýšení relevance vyhledaných dokumentů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Navržený systém bude testován na vícejazyčných korpusech vytvořených v rámci plnění cíle 2006-22 a 2007-19.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Systém bude testován na relevanci získávaných výsledků jak pro české a anglické prostředí, tak i při křížovém zpracování. Provedeno bude také srovnání výsledků s přístupem aplikovaným ve vyhledávači Google.

Číslo aktivity

17

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity
Rozšiřování uživatelského dotazu
Zahájení aktivity
2.1.2007
Ukončení aktivity
30.12.2007
Popis aktivity
V rámci projektu jsme navrhli prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jednou s možností vylepšující relevanci dokumentů vyhledávaných uživatelem je rozšíření dotazu. Touto progresivní oblastí se hodláme zabývat a očekáváme zlepšení relevance a pokrytí výsledků hledání.
Plánované indikátory dosažení - očekávané výsledky aktivity
Navržený systém bude obsahovat možnost rozšiřování uživatelských dotazů s použitím tezauru EWN.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Systém bude testován na relevanci získávaných výsledků jak pro české a anglické prostředí, tak i při křížovém zpracování. Porovnání bude provedeno ručně zaškolenými evaluátory.

Číslo aktivity
18
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Metody disambiguace ve vyhledávací úloze
Zahájení aktivity
2.1.2007
Ukončení aktivity
1.9.2007
Popis aktivity
Rozlišení významů slova je nezbytným krokem pro většinu aplikací zpracovávajících přirozený jazyk. Jedná se o klíčovou úlohu pro správné porozumění sdělení, uplatňuje se v komunikaci člověk-počítač. Disambiguační modul hodláme začlenit do systému vyhledávání.
Plánované indikátory dosažení - očekávané výsledky aktivity
Funkční aplikační modul schopný provádět disambiguaci slov v anglickém a českém jazyce v souvislosti s vyhledávací úlohou.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Výsledky systému budou testovány na vybraných dokumentech a porovnány s očekávaným výsledkem. Evaluace bude prováděna zaškolenými pracovníky.

Číslo aktivity
19
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Klasifikační metody v multijazykovém prostředí
Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Vzhledem k rostoucímu významu mezinárodní kooperace nabývá na důležitosti zpracování dokumentů v různých jazycích. Proto jsme se rozhodli klasifikační metody rozšířit a modifikovat pro použití v multilingválním prostředí. V navazujícím výzkumu se chystáme ověřit různé lemmatizační metody a jejich vliv na klasifikační úlohu.

Plánované indikátory dosažení - očekávané výsledky aktivity

Provedeme klasifikaci dokumentů lemmatizovaných různými metodami a porovnáme jejich vliv na kvalitu výsledných korpusů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ověření vlivu provedeme na klasifikační úloze. Použijeme klasifikátor multinomial Naive bayes a srovnáme výsledky s ostatními přístupy.

Číslo aktivity

20

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Doplnění jazykových korpusů o další evropské jazyky

Zahájení aktivity

2.1.2007

Ukončení aktivity

1.7.2007

Popis aktivity

Vzhledem k rozšiřující se mezinárodní kooperaci nabývá na důležitosti zpracování dokumentů v různých jazycích. Proto jsme se rozhodli datové korpusy rozšířit o další evropské jazyky.

Plánované indikátory dosažení - očekávané výsledky aktivity

Očekáváme vytvoření rozsáhlé multilinguální databáze obsahující dokumenty ve významnějších evropských jazycích. Na těchto korpusech lze následně provádět testování vytvářených algoritmů a navrhovaných postupů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

U vytvořeného korpusu zjistíme standardní metriky – konkrétně počty slov, dokumentů, distribuci délky dokumentů atp.

Číslo aktivity

21

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření rozhraní ke klasifikátoru SVM

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.6.2007

Popis aktivity

Klasifikátor SVM je v současné době jediný klasifikátor, který dosahuje vynikajících výsledků prakticky ve všech oblastech zpracování dat, včetně zpracování textu. Stalo se proto jednou z priorit tohoto projektu využít jeho vlastností a prozkoumat jeho efektivitu – například při použití různých modelů charakterizujících textové dokumenty. Protože je již na internetu k dispozici jeho univerzální implementace (viz http://svmlight.joachims.org),

bylo v rámci této aktivity vytvořeno základní aplikační rozhraní, které umožňuje použití různých technik využívaných při zpracování textu současně s klasifikátorem SVM. Účelem rozhraní je zjednodušit použití klasifikátoru SVM pro potřeby zpracování textu a odstranit některé kroky, které musejí být díky snaze o co jeho nejuniverzálnější použití v různých oblastech prová-děny. V minulém období, ve kterém již tato aktivita byla řešena, byl vytvořen funkční základ potřebný pro implementaci dalších nadstaveb. Díky tomu bude dále možné do rozhraní zakomponovat možnost použít k jednotlivým slovům reprezentujícím textový dokument také bigramy a 2-itemsety. Tímto obohacením modelu textového dokumentu očekáváme obecné zlepšení výsledků klasifikace, což může být prokázáno díky již implementovaným micro-F1 a macro-F1 mírám, které jsou obecně uznávány a výsledky experimentů tudíž mohou být porovnány i s jinými výzkumnými pracemi. V minulém období byl také objeven úspěšnější přístup k převodu reprezentace textového dokumentu do podoby akceptované SVM klasifikátorem, kterým v nadcházejícím období plánujeme nahradit stávající přístup.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity

22

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Implementace klasifikační metody Naive Bayes rozšířené o současné zpracování bigramů a 2-itemsetů

Zahájení aktivity

1.7.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Vlastnosti metody Naive Bayes a její úspěšnost při klasifikaci textových dokumentů byly poměrně podrobně prozkoumány. Stále se zde však objevují nové prostory pro zkoušení nových přístupů vedoucích ke zlepšení dosahovaných výsledků. Jedním z nich je i možnost obohatit obecně používaný bag-of-words model dokumentu (tedy model založený na jen jednotlivých slovech) o další položky. Těmi mohou být například slovní n-gramy nebo itemsety. Doposud byly n-gramy a itemsety zkoumány vždy odděleně, naším cílem se však stalo nejen ověřit, které z obou přístupů poskytují lepší výsledky, ale navíc se i pokusit oba přístupy k zlepšení úspěšnosti klasifikace zkombinovat za účelem vylepšení dosavadních výsledků. A právě uskutečněním této aktivity se nám otvírá možnost tyto experimenty realizovat. Částečně zde byly využity poznatky a principy použité ve výše popsané aktivitě b), zejména se jednalo o fázi vyhodnocení celkové úspěšnosti klasifikace. Díky dokončení této aktivity jsme již měli možnost dosáhnout poměrně zajímavých výsledků, které jsme úspěšně publikovali – viz odstavec 4.2.1. V následujícím období plánujeme použít bigramy a 2-itemsety současně k obohacení modelů textových dokumentů a ověřit přínos tohoto přístupu na standardizovaných kolekcích, podobně jako tomu bylo v naší práci dokončené v předchozím období. Naším záměrem je také pokusit se ověřit tento přístup na textových kolekcích v českém jazyce, pro který v současné době standardizované kolekce neexistují. Plánujeme proto vytvořit vlastní testovací kolekci z dat ČTK (České tiskové kanceláře).

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy, možnost různých nastavení nutných pro experimenty.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity

23

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření modulu pro zpracování a vyhodnocení dat získaných z Internetu

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Součástí připravovaného systému pro procházení webu a automatickou identifikaci internetových stránek podle tématu, o kterém pojednávají, bylo v první fázi projektu vytvoření aplikace označované jako webový spider. Jejím cílem je zadanou výchozí stránku zpracovat, získat z ní nebo určit předem definované údaje včetně odkazů na další stránky, které budou následně stejným způsobem zpracovány. Tento cíl se v uplynulém období podařilo splnit. Důležitou vlastností této aplikace je především její modularita, která jednoduchým způsobem dovoluje snadnou modifikaci. Možné využití tedy nepředstavuje jen automatické vytváření námětových korpusů, ale ve spojení s dalšími vhodnými moduly představuje vhodný prostředek pro ověření navržených algoritmů určených přímo k práci s webovým prostředím. Ve spojení s vhodným analyzátorem dat získaných z webového spidera je možné kompletně mapovat určitou tématickou doménu, v našem případě například servery obsahující závadné (protizákonné) materiály, které se vykytují v rámci určitého území – například České republiky nebo států Evropské unie. A právě tento analyzátor plánujeme v nadcházejícím období navrhnout a implementovat. Neocenitelnými údaji poskytovanými tímto analyzátorem budou bezpochyby statistiky počtu výskytů určitých slov na závadných webových stránkách, údaje o často se zde vyskytujících emailových adresách, analýza významnosti jednotlivých webových serverů z pohledu množství závadných dat, a podobně.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - modulárně členěná, možnost různých nastavení.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, nastavení konfiguračních parametrů a testování běhu aplikace, zakomponování nově vytvořeného modulu, kontrolní úprava funkčnosti stávajících modulů

Číslo aktivity

24

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Výběr a implementace algoritmů pro ohodnocení a výběr charakteristických položek

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Pro obohacení modelů textových dokumentů v současné době testujeme použití n-gramů a itemsetů (dále jen položek). Základním problémem je ovšem nejen jejich generování z textu, ale i výběr jen těch charakteristických položek, které nejvíce přispějí k zvýšení úspěšnosti klasifikace. K tomuto účelu již byly navrženy různé přístupy. Po prostudování dostupné literatury a výběru jen těch přístupů, které obecně poskytují výborné výsledky, přejdeme k

jejich implementaci s ohledem na zpracování co největšího množství textových dat. V současné době sice máme k dispozici pro experimentální účely několik vybraných metod, nicméně pro zpracování statisticky významného počtu dokumentů je potřeba stávající řešení přepracovat a výrazně doplnit.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - modulárně členěná, možnost různých relevantních nastavení.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, nastavení konfiguračních parametrů a testování poskytovaných výstupů.

Číslo aktivity

25

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Lexikální databáze Verbalex obsahující valenční rámce českých sloves a jejich vazby na Princetonský WordNet v.2.0

Zahájení aktivity

1.7.2006

Ukončení aktivity

Popis aktivity

Kompletování valenčních rámců českých sloves do synsetů, doplňování definic a přiřazování překladových ekvivalentů z Princetonského WordNetu 2.0. Jde primárně o pracné manuální přiřazování vyžadující kvalifikované pracovníky.

Plánované indikátory dosažení - očekávané výsledky aktivity

Zpracované rámce budou k dispozici prostřednictvím stejnojmenného webového rozhraní. Práce pravděpodobně nebudou ukončeny ke konci r.2007.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Počítá se s publikováním zmíněných výsledků a využitím v aplikacích.

Číslo aktivity

26

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Korpus syntaktických stromů včetně morfologické desambiguace

Zahájení aktivity

1.1.2007

Ukončení aktivity

Popis aktivity

Příprava korpusu syntaktických stromů včetně plné morfologické desambiguace.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledný korpus

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Korpus dostupný prostřednictvím korpusového manažeru.

Číslo aktivity

27

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Korpus vzorových přepisů vět do konstrukcí TILu a přiřazení odpovídajících sémantických reprezentací

Zahájení aktivity

1.1.2007

Ukončení aktivity

Popis aktivity

Tvorba korpusu vzorových (typových) přepisů českých (anglických) vět do konstrukcí TILu spolu s přiřazením odpovídajících sémantických reprezentací.

Plánované indikátory dosažení - očekávané výsledky aktivity

Vytvořený korpus

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vytvořený korpus bude dostupný přes vhodný korpusový nástroj (manažer).

Číslo aktivity

28

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Návrh a implementace guesseru - modulu pro automatické doplňování morfologické databáze češtiny

Zahájení aktivity

1.1.2007

Ukončení aktivity

31.12.2008

Popis aktivity

Modul automaticky doplňující nová slova do morfologické databáze češtiny je nutný pro realistický provoz ostatních komponent vytvářených v projektů - bude navržena jeho struktura a podle možnosti implementována. Vzhledem ke své obtížnosti nebude pravděpodobně aktivita ukončena v r. 2007.

Plánované indikátory dosažení - očekávané výsledky aktivity

Vlastní guesser

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Předpokládáme, že výsledek bude zpracován formou publikace o modulu (včetně evaluace).

Číslo aktivity

29

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Detekce plagiátů (spamů) s využitím sémantických znalostí

Zahájení aktivity

1.1.2007

Ukončení aktivity

1.12.2009

Popis aktivity

Porovnání existujících koncepcí pro zjišťování plagiátů a příprava návrhu algoritmu, který bude schopen pracovat

se znalostmi sémantické povahy. Aktivita bude pokračovat v r. 2007-2009.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledky získané porovnáním, jejich analýza s důsledky pro budování inteligentního rozpoznávače plagiátů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Předpokládáme, že bude připravena publikace obsahující evaluaci.

Číslo aktivity

30

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Rozpoznávání anaforických vztahů ve volných textech

Zahájení aktivity

1.7.2006

Ukončení aktivity

31.12.2009

Popis aktivity

Primárně bude věnována pozornost anaforickým vztahům pronominálním - budou testovány existující algoritmy pro rozpoznávání anaforických vztahů ve volném textu, posouzena jejich vhodnost pro češtinu. Budou se řešit vazby na moduly, které jsou pro rozpoznávání anaforických vztahů nezbytné, konkrétně na syntaktický analyzátor synt - v této souvislosti bude potřeba navrhnout vhodné formáty a notační konvence.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výběr úspěšného algoritmu pro rozpoznávání anaforických vztahů, podle potřeby jeho modifikace. Pokuus o implementaci.

Vzhledem k obtížnosti dané problematiky bude aktivita pokračovat i v r. 2008 a 2009.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Předpokládáme publikační výstupy a experimenty ověřující úspěšnost řešení, které bude zvoleno.

Číslo aktivity

31

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Analýza problematiky vytváření grafiky a webovských presentací prostřednictvím dialogových systémů.

Zahájení aktivity

1.1.2007

Ukončení aktivity

Popis aktivity

V oblasti analýzy vytváření webovských stránek a počítačové grafiky v kontextu sémantického webu aplikacemi v asistivních technologiích bude vytvořena taxonomie webovských presentací a na jejím základě budou implementovány rámce pro formalizaci na bázi klasifikačních systémů. Dále budou vytvořeny základní grafické ontologie zaměřené na na zpřístupňování informací a možnosti vytváření internetových prezentací pro nevidomé zohledňující požadavky přístupnosti vůči nevidomým (internetový standard Web Content Accessibility). Na formální bázi klasifikačních systémů popisujících elementární ontologie, který do formátu SVG integruje popis odpovídajícího grafického objektu, odvozený z historie jeho dialogového vygenerování, budou vytvořeny a implementovány dialogové strategie umožňující vytvářet grafiku, jejíž popis je čitelně zahrnut do formátu grafického objektu takovým způsobem, že jej lze využívat k automatickému získání popisu vygenerované grafiky na webovských presentacích, které se tak stávají přístupnější pro nevidomé uživatele.

Plánované indikátory dosažení - očekávané výsledky aktivity

publikace

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

---

Číslo aktivity

32

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření korpusu matematických textů pro vyhledávání na webu

Zahájení aktivity

1.1.2007

Ukončení aktivity

31.12.2008

Popis aktivity

Bude vytvořen korpus více než 100000 stran matematických textů a budou studovány možnosti indexování a vyhledávání strukturních informací (matematiky) ve formátech TeX a MathML. Na to naváže plánovaná aktivita v roce 2008 a 2009 zahrnující klasifikaci a strojové učení klasifikace matematických textů dle AMS 2000 classification scheme.

Plánované indikátory dosažení - očekávané výsledky aktivity

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

---

3.2. NÁVRH ZMĚN V ŘEŠENÍ PROJEKTU - rok 2007

Pč.	Typ	Popis
1	návrh změn v řešení projektu	Při řešení projektu v roce 2007 nejsou plánovány žádné závažné změny řešitelského plánu.

3.3. NÁVRH ZMĚN V NÁKLADECH - rok 2007

Pč.	Typ	Popis
1	návrh změn v nákladech	Kromě čerpání nevyužitých finančních prostředků z roku 2006 (převedených do fondu účelově určených prostředků) na dohody o pracích (zejména pro studenty spolupracující na tvorbě korpusů) a podporu aktivit nově přijatých doktorandů nejsou žádné další změny v nákladech na řešení projektu plánovány.

4. PŘÍLOHY

4.1. ZPRÁVA O POSTUPU ŘEŠENÍ PROJEKTU - rok 2006

4.1.1. POPIS ŘEŠENÍ PROJEKTU - seznam

	Soubor	
	Zprava_2C06009_odst411.doc	
	Zprava_2C06009_Brno.doc	

4.1.2. DOSAŽENÉ VÝSLEDKY

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/01/2006**

Název výsledku

Andrš, D., Ekštein, K.: Koartikulační kompozitní modely v akusticko-fonetickém dekódování. In: Kognice a umělý život VI, Ediční středisko FPF SU, Opava, 2006.

Abstrakt

Jedním z klíčových problémů strojového porozumění přirozené řeči je sémantická analýza, tj. postup, kterým inteligentní stroj z výpovědi v přirozeném jazyce „dobývá“ její význam neboli buduje instanci formalismu, která umožňuje zaznamenat a dále zpracovávat relevantní informace v promluvě obsažené. Tento proces se při současném stavu poznání zatím ani zdaleka nepřibližuje optimálnímu. Jednak není k dispozici vhodný univerzální formalismus (jen několik problémově orientovaných přiblížení) a jednak existují jen velmi nesystematické a neobecné metody vytváření jeho instancí. Mezinárodní umělý jazyk esperanto, vytvořený v 80. letech 19. století polským lékařem L. L. Zamenhofem, se překvapivě nabízí jako jeden z možných formalismů, vhodných k zachycení jak syntaktické, tak sémantické informace obsažené v promluvě v přirozené řeči. Tento článek uvádí některé teoretické úvahy týkající se extrakce a ukládání sémantické informace ve formě jazyka, který je současně formální i přirozený, výhody a omezení takového přístupu a příklady, jak by mohlo esperanto nahradit některé v současnosti využívané formalismy.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Andrš, D., Ekštein, K.: Koartikulační kompozitní modely v akusticko-fonetickém dekódování. In: Kognice a umělý život VI, Ediční středisko FPF SU, Opava, 2006. ISBN 80-7248-355-2.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/02/2006**

Název výsledku

Ekštein, K., Andrš, D.: Esperanto jako meta-jazyk analýzy přirozené řeči. In: Kognice a umělý život VI, Ediční středisko FPF SU, Opava, 2006. ISBN 80-7248-355-2

Abstrakt

Jedním z klíčových problémů strojového porozumění přirozené řeči je sémantická analýza, tj. postup, kterým inteligentní stroj z výpovědi v přirozeném jazyce „dobývá“ její význam neboli buduje instanci formalismu, která umožňuje zaznamenat a dále zpracovávat relevantní informace v promluvě obsažené. Tento proces se při současném stavu poznání zatím ani zdaleka nepřibližuje optimálnímu. Jednak není k dispozici vhodný univerzální formalismus (jen několik problémově orientovaných přiblížení) a jednak existují jen velmi nesystematické a neobecné metody vytváření jeho instancí. Mezinárodní umělý jazyk esperanto, vytvořený v 80. letech 19. století polským lékařem L. L. Zamenhofem, se překvapivě nabízí jako jeden z možných formalismů, vhodných k zachycení jak syntaktické, tak sémantické informace obsažené v promluvě v přirozené řeči. Tento článek uvádí některé teoretické úvahy týkající se extrakce a ukládání sémantické informace ve formě jazyka, který je současně formální i přirozený, výhody a omezení takového přístupu a příklady, jak by mohlo esperanto nahradit některé v současnosti využívané formalismy.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
02	Ekštein, K., Andrš, D.: Esperanto jako meta-jazyk analýzy přirozené řeči. In: Kognice a umělý život VI, Ediční středisko FPF SU, Opava, 2006. ISBN 80-7248-355-2.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/03/2006**

Název výsledku

Fiala D., Tesař R., Ježek K., Rousselot F.: “Extracting Information from Web Content and Structure”. The 9th International Conference on Information Systems Imp

Abstrakt

Web is a vast data repository. By mining from this data efficiently, we can gain valuable knowledge. Unfortunately, in addition to useful content there are also many Web documents considered harmful (e.g. pornography, terrorism, illegal drugs). Web mining that includes three main areas – content, structure, and usage mining – may help us detect and eliminate these sites. In this paper, we concentrate on applications of Web content and Web structure mining. First, we introduce a system for detection of pornographic textual Web pages. We discuss its classification methods and depict its architecture. Second, we present analysis of relations among Czech academic computer science Web sites. We give an overview of ranking algorithms and determine importance of the sites we analyzed.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
03	Fiala D., Tesař R., Ježek K., Rousselot F.: “Extracting Information from Web Content and Structure”. The 9th International Conference on Information Systems Implemen-tation and Modelling (ISIM ’06), Přerov, Czech Republic, ISBN 80-86840-19-0, pages 133-140, CEUR-WS proceedings, Vol. 180, ISSN 1613-0073, http://ceur-ws.org/Vol-180, 2006.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/04/2006**

Název výsledku

Fiala D., Jezek, K., Rousellot, F.: Finding Autoritative Researchers on Academic Web Sites, Enformatica, Vol. 17, Dec. 2006, pp. 74 – 79, ISSN 1305 - 5313

Abstrakt

In this paper, we present a methodology for finding authoritative researchers by analyzing academic Web sites. We show a case study in which we concentrate on a set of Czech computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones using several common ranking algorithms. We then examine the contents of the research papers present on these sites and determine the most authoritative Czech authors.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
04	Fiala D., Jezek, K., Rousellot, F.: Finding Autoritative Researchers on Academic Web Sites, Enformatica, Vol. 17, Dec. 2006, pp. 74 – 79, ISSN 1305 - 5313	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/05/2006**

Název výsledku

Konopík, M.: Stochastic Semantic Analysis, PhD Workshop 2006, Hrubá Skála, Czech Republic, 2006

Abstrakt

Speech is the most natural way of human communication. Therefore, there is an effort to incorporate speech control into human-computer interfaces. However, nobody likes the idea of remembering a large amount of specific commands and so the ability of understanding the meaning of an utterance is crucial for many speech-enabled computer systems. This article describes a very promising method suitable for analysis of the meaning contained in an utterance. The described method extends Markov models so that every Markov state stores a semantic vector. This extension allows the model to capture hierarchical structures.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
05	Konopík, M.: Stochastic Semantic Analysis, PhD Workshop 2006, Hrubá Skála, Czech Republic, 2006	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG
05	Konopík, M.: Stochastic Semantic Analysis, PhD Workshop 2006, Hrubá Skála, Czech Republic, 2006	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/06/2006**

Název výsledku

Konopík, M.: Stochastic Semantic Parsing, Technical Report No. DCSE/TR-2006-01, KIV ZČU Plzeň, 2006.

Abstrakt

Speech is the most natural way of human communication. Therefore, there is an effort to incorporate speech control into human-computer interfaces. However, nobody likes the idea of remembering a large amount of specific commands and so the ability of understanding the meaning of an utterance is crucial for many speech-enabled computer systems. This work describes a very promising method suitable for analysis and parsing of the meaning contained in an utterance.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
06	Konopík, M.: Stochastic Semantic Parsing, Technical Report No. DCSE/TR-2006-01, KIV ZČU Plzeň, 2006.	V - Oponovaná výzkumná zpráva určená pro státní správu	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/07/2006**

Název výsledku

Kral, P., Cerisasa, C., and Kleckova, J.: Automatic Dialog Acts Recognition based on Sentence Structure. In: ICASSP '06 Proceedings, Toulouse, France, 2006, pp.

Abstrakt

This paper deals with automatic dialog acts (DAs) recognition in Czech. Our work focuses on two applications: a multimodal reservation system and an animated talking head for hearing-impaired people. In that context, we consider the following DAs: statements, orders, investigation questions and other questions. The main goal of this paper is to propose, implement and evaluate new approaches to automatic DAs recognition based on sentence structure and prosody. Our system is tested on a Czech corpus that simulates a task of train tickets reservation. With lexical-only information, the classification accuracy is 91 %. We proposed two methods to include sentence structure information, which respectively give 94 % and 95 %. When prosodic information is further considered, the recognition accuracy reaches 96 %.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
07	Kral, P., Cerisasa, C., and Kleckova, J.: Automatic Dialog Acts Recognition based on Sentence Structure. In: ICASSP '06 Proceedings, Toulouse, France, 2006, pp. 61-64.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/08/2006**

Název výsledku

Kral, P., Cerisara, C., Kleckova, J., Pavelka, T.: Sentence Structure for Dialog Act Recognition in Czech. Proceedings of ICTTA´06, Damascus, Syria, 2006

Abstrakt

This paper deals with automatic dialog acts (DAs) recognition in Czech based on sentence structure. We consider the following DAs: statements, orders, yes/no questions and other questions. In our previous works, we have proposed, implemented and evaluated new approaches to automatic DAs recognition based on sentence structure and prosody. The word sequences were manually transcribed. The main goal of this paper is to evaluate the performances of our approaches when these word sequences are unknown and estimated from a speech recognizer. Our system is tested on a Czech corpus that simulates a task of train tickets reservation. When manual transcription is used, classi□cation accuracy without and with sentence structure models is 91 %, 94 % and 95 %. The recognition accuracy reaches 96 % with prosodic combination. When word sequences are estimated from a speech recognizer, the classification score is 88 % without and 91 % and 92 % with sentence structure models. The combination with prosody gives 93 % of accuracy.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
08	Kral, P., Cerisara, C., Kleckova, J., Pavelka, T.: Sentence Structure for Dialog Act Recognition in Czech. Proceedings of ICTTA´06, Damascus, Syria, 2006	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/09/2006**

Název výsledku

Kral, P., Kleckova, J., and Cerisasa, C.: Automatic Dialog Acts Recognition based on Words Clusters. In WESPAC IX 2006, Seoul, Korea, 2006.

Abstrakt

This paper deals with automatic dialog acts (DAs) recognition in Czech. A Dialog act is defined by J. L. Austin [1] as a meaning of an utterance at the level of illocutionary force. The four following DAs are considered: statements, orders, yes/no questions and other questions. In our previous works, we proposed, implemented and evaluated two new approaches to automatic DAs recognition based on sentence structure. These methods have been validated on a Czech corpus that simulates a task of train tickets reservation. The main goal of this paper is to propose a new approach to solve the problem of lack of training data for automatic DA recognition. This approach clusters the words in the sentence into several groups using maximization of mutual information between two neighbor word classes. The classification accuracy of the unigram model (our baseline approach) is 91 %. The proposed method, a clustered unigram model, reduces the DA error rate by 12 %.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
09	Kral, P., Kleckova, J., and Cerisasa, C.: Automatic Dialog Acts Recognition based on Words Clusters. In WESPAC IX 2006, Seoul, Korea, 2006.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/10/2006**

Název výsledku

Matoušek, V., Nestorovič, T.: Návrh hlasové komunikace s navigačním systémem automobilu a její implementace v jazyce VoiceXML. In: Sborník referátů mezinárodní

Abstrakt

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
10	Matoušek, V., Nestorovič, T.: Návrh hlasové komunikace s navigačním systémem automobilu a její implementace v jazyce VoiceXML. In: Sborník referátů mezinárodní konference NavAge´06, Telematix Praha, 2006	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/11/2006**

Název výsledku

Matoušek, V., Nestorovič, T.: Entwurf der Sprachkommunikation mit einem Car-Navigationssystem und ihre Implementation in der VoiceXML Sprache. In: Elektronisch

Abstrakt

Ein experimentelles Navigationssystem wurde an der Universität in Pilsen entwickelt, das mit dem Autofahrer nur über Sprache kommuniziert. Erst mussten einige strenge Voraus setzungen des Systemansatzes definiert und erfüllt werden – so soll das System über die GPS-Satellitenanlage mit einer Verkehrsinformationen-Datenbank kommunizieren und diese Anlage zur Bestimmung der Wagenposition verwenden, es sollte auf Basis des "Standardrechners" entwickelt werden und in diesem Zusammenhang wird es mit einer robusten und zuverlässigen Spracherkennung sowie auch mit Qualitäts-Sprachausgabe ausgerüstet. Es werden die möglichen Alternativen der Dialogführung diskutiert, Entwicklung eines experimentellen und auf das professionelle Navigationssystem bezogenen Dialogs kurz beschrieben, der vereinfachte Entwurf von Basisfunktionen des Systems schrittweise ausgeführt und die Implementierung einiger Systemfunktionen in der VoiceXML Programmiersprache präsentiert.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
11	Matoušek, V., Nestorovič, T.: Entwurf der Sprachkommunikation mit einem Car-Navigationssystem und ihre Implementation in der VoiceXML Sprache. In: Elektronische Sprachsignalverarbeitung, Tagungsband der 17. Konferenz, Freiberg, August 2006; TUD Press, 2006	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	NEM

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/12/2006**

Název výsledku

Matoušek, V., Michalicová, J., Mouček, R.: Czech Explanatory Dictionary and its Computer Implementation. In: Sborník konference Corpora 2006, University of St.

Abstrakt

The article deals with the overview and splitting of the dictionaries used in several language engineering applications, analysis of the contemporary attempt to the creation of on-line explanatory language dictionaries, analysis of the content of keyword paragraphs of explanatory dialectic language dictionary, and keynote issues of the implementation of this kinds of dictionary.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
12	Matoušek, V., Michalicová, J., Mouček, R.: Czech Explanatory Dictionary and its Computer Implementation. In: Sborník konference Corpora 2006, University of St. Petersburg, October 2006	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/13/2006**

Název výsledku

Mouček, R.: Sémantická a epizodická paměť ve vztahu k počítačovému zpracování sémantiky přirozeného jazyka. In: Kognice a umělý život VI, Ediční středisko FPF S

Abstrakt

Dlouhodobá vědomá paměť a její složky pro zapamatování událostí vázaných na kontext, prostor a čas - epizodická paměť a zapamatování faktů, pojmů a významů "kontextově nezávislých" - sémantická paměť mají zajímavou spojitost s moderními metodami a modely návrhu softwarových produktů a výpočetních systémů. Tuto spojitost lze nalézt nejen při modelování specifického softwaru v oblastech umělé inteligence a zpracování sémantiky přirozeného jazyka, ale i při modelování běžných softwarových produktů. Pozornost je pak věnována možnostem počítačového zpracování sémantiky přirozeného jazyka (sémantické analýze a interpretaci).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
13	Mouček, R.: Sémantická a epizodická paměť ve vztahu k počítačovému zpracování sémantiky přirozeného jazyka. In: Kognice a umělý život VI, Ediční středisko FPF SU, Opava, 2006. ISBN 80-7248-355-2.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/14/2006**

Název výsledku

Mouček, R.: Natural Language Semantics and Problem of Layers, PhD Workshop 2006, Hrubá Skála, Czech Republic

Abstrakt

This paper deals with one of the essential problems connected with processing of natural language semantics – an effort to divide natural language semantics into several layers and process them separately. We introduce some existing solutions according to field of science dealing with natural language processing. Then we discuss the advantages and disadvantages of this approach in contrast to processing of semantics as the indivisible unit. As a result we describe a contemporary approach of our research group to the problem of natural language processing.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
14	Mouček, R.: Natural Language Semantics and Problem of Layers, PhD Workshop 2006, Hrubá Skála, Czech Republic	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/15/2006**

Název výsledku

Pavelka, T.: LDec: One Pass Time Synchronous Decoder, PhD Workshop 2006, Hrubá Skála, Czech Republic

Abstrakt

The search for the most probable word sequence is in automatic speech recognition called decoding and is usually carried out by the Viterbi algorithm, an efficient search strategy based on dynamic programming. The paper discusses implementation issues and methods to further reduce the computational costs when performing recognition with large vocabularies and stochastic language models. Such methods can be divided into two categories: those that eliminate unnecessary computations (such as tree structured lexicons) and those that exclude computations according to the probability that those computations will lead to the desired result (pruning).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
15	Pavelka, T.: LDec: One Pass Time Synchronous Decoder, PhD Workshop 2006, Hrubá Skála, Czech Republic	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/16/2006**

Název výsledku

Steinberger, J., Ježek, K.: Sentence Compression for the LSA-based Summarizer. Proceedings of the 7th International Conference on Information Systems Implementa

Abstrakt

We present a simple sentence compression approach for our summarizer based on latent semantic analysis (LSA). The summarization method assesses each sentence by an LSA score. The compression algorithm removes unimportant clauses from a full sentence. Firstly, a sentence is divided into clauses by Charniak parser, then compression candidates are generated and finally, the best candidate is selected to represent the sentence. The candidates gain an importance score which is directly proportional to its LSA score and indirectly to its length. We evaluated the approach in two ways. By intrinsic evaluation we found that the compressions produced by our algorithm are better than baseline ones but still worse than what humans can make. Then we compared the resulting summaries with human abstracts by a standard n-gram based ROUGE measure.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
16	Steinberger, J., Ježek, K.: Sentence Compression for the LSA-based Summarizer. Proceedings of the 7th International Conference on Information Systems Implementation and Modelling, pp. 141-148, MARQ Ostrava, Přerov, Czech Republic, 2006, ISBN 80-86840-19-0.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/17/2006**

Název výsledku

Tatarnikova, M., Tampel, I., Oparin, I., Khokhlov, Y.: Building Acoustic Models for a Large Vocabulary Continuous Speech Recognizer for Russian. In: Proceedings

Abstrakt

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
17	Tatarnikova, M., Tampel, I., Oparin, I., Khokhlov, Y.: Building Acoustic Models for a Large Vocabulary Continuous Speech Recognizer for Russian. In: Proceedings of XI. International Conference on Speech and Computer SPECOM'06, St.Petersburg, Russia, 2006. pp. 83-87.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/18/2006**

Název výsledku

Tesar R., Poesio M., Strnad V., Jezek K.: “Extending the Single Words-Based Document Model: A Comparison of Bigrams and 2-Itemsets”.

Abstrakt

The basic approach in text categorization is to represent documents by single words. However, often other features are utilized to achieve better classification results. In this paper, our attention is focused on bigrams and 2-itemsets. We compare the performance improvement in terms of classification accuracy when these features are used to extend the single words-based document representation on two standard text corpora: Reuters-21578 and 20 Newsgroups. For this comparison we use the multinomial Naive Bayes classifier and five different feature selection approaches. Algorithms for bigrams and 2-itemsets discovery are presented as well. Our results show a statistically significant improvement when bigrams and also 2-itemsets are incorporated. However, in the case of 2-itemsets it is important to use an appropriate feature selection method. On the other hand, even when a simple feature selection approach is applied to discover bigrams the classification accuracy improves. The conclusion is that, in our case, it is not very effective to extend document representation with 2-itemsets because bigrams achieve better results and discovering them is less resource-consuming.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
18	Tesar R., Poesio M., Strnad V., Jezek K.: “Extending the Single Words-Based Document Model: A Comparison of Bigrams and 2-Itemsets”. The 2006 ACM Symposium on Document Engineering (DocEng’06), Amsterdam, Netherlands, ACM press, ISBN 1-59593-515-0, pp. 138–146, http://doi.acm.org/10.1145/1166160.1166197, October 2006.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/19/2006**

Název výsledku

Toman M., Tesar R., Jezek K.: “Influence of Word Normalization on Text Classification”. The 1st International Conference on Multidisciplinary Information Scienc

Abstrakt

In this paper we focus our attention on the comparison of various lemmatization and stemming algorithms, which are often used in nature language processing (NLP). Sometimes these two techniques are considered to be identical, but there is an important difference. Lemmatization is generally more utilizable, because it produces the basic word form which is required in many application areas (i.e. cross-language processing and machine translation). However, lemmatization is a difficult task - especially for highly inflected natural languages having a lot of words for the same normalized word form. We present a novel lemmatization algorithm that utilizes the multilingual semantic thesaurus Eurowordnet (EWN). We describe the algorithm in detail and compare it with other widely used algorithms for word normalization on two different corpora. We present promising results obtained by our EWN-based lemmatization approach in comparison to other techniques. We also discuss the influence of the word normalization on classification task in general. In overall, the performance of our method is good and it achieves similar precision and recall in comparison with other word normalization methods. However, our experiments indicate that word normalization does not affect the text classification task positively.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
00	Toman M., Tesar R., Jezek K.: “Influence of Word Normalization on Text Classification”. The 1st International Conference on Multidisciplinary Information Sciences & Technologies (InSciT 2006), Merida, Spain, ISBN 84-611-3105-3, pp. 354-358, October 2006.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/20/2006**

Název výsledku

Toman, M.; Steinberger J.: Jezek K. Searching and Summarizing in a Multilingual Environment, ELPUB2006 – Proceedings of the 10th International Conference on Ele

Abstrakt

Multilingual aspects have been gaining more and more attention in recent years. This trend has been accentuated by the global integration of European states and the vanishing cultural and social boundaries. The ever increasing use of foreign languages is due to the information boom caused by the emergence of easy internet access. Multilingual text processing has become an important field bringing a lot of new and interesting problems. Their possible solutions are proposed in this paper. Its first part is devoted to methods for multilingual searching, the second part deals with the summarization of retrieved texts. We tested several novel processing techniques: a languageindependent storage format, semantic-based indexing, query expansion or text summarization leading to faster and easier retrieval and understanding of documents. We implemented a prototype system named MUSE (Multilingual Searching and Extraction) and compared its qualities with the state-ofthe-art search engine – Google. The results seem to be promising; MUSE shows high correlation with the market-leading products. Although for our experiments we used Czech and English articles, the main principle applies to other languages as well.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
20	Toman, M.; Steinberger J.: Jezek K. Searching and Summarizing in a Multilingual Environment, ELPUB2006 – Proceedings of the 10th International Conference on Electronic Publishing, Bansko, Bulgaria, 2006, ISBN 978-954-16-0040-5, 2006, pp. 257-266.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/21/2006**

Název výsledku

Pomikálek, J. Řehůřek, R., The Influence of Preprocessing Parameters on Text Categorization, paper accepted at XIX Int. Conference on Computer and Systems Scien

Abstrakt

Článek obsahuje výchozí výsledky týkající se kategorizace textů.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
21	Pomikálek, J. Řehůřek, R., The Influence of Preprocessing Parameters on Text Categorization, paper accepted at XIX Int. Conference on Computer and Systems Science and Egineering, January 29-31, Bangkok 2007, (in print).	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/22/2006**

Název výsledku

Kopeček, I., Ošlejšek, R., Creating Pictures by Dialogue. In Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006. Berlin: Spr

Abstrakt

Text se věnuje analýze vytváření webovských stránek a počítačové grafiky v kontextu sémantického webu a s aplikacemi v asistivních technologiích, zejména s ohledem na zpřístupňování informací a možnosti vytváření internetových prezentací pro nevidomé.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
22	Kopeček, Ivan - Ošlejšek, Radek. Creating Pictures by Dialogue. In Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006. Berlin : Springer-Verlag, 2006,. od s. 61-68, 8 s. ISBN 3-540-36020-4.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/23/2006**

Název výsledku

Kopeček, I., Ošlejšek, R., The Blind and Creating Computer Graphics. In Proceedings of the Second IASTED International Conference on Computational Intelligence.

Abstrakt

V článku se probírá základní koncept přístupu, který současně splňuje a zajišťuje požadavky přístupnosti vůči nevidomým (internetový standard Web Content Accessibility).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
23	Kopeček, Ivan - Ošlejšek, Radek. The Blind and Creating Computer Graphics. In Proceedings of the Second IASTED International Conference on Computational Intelligence. Anaheim, Calgary, Zurich : ACTA Press, 2006, od s. 343-348, 6 s. ISBN 0-88986-602-3.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/24/2006**

Název výsledku

Kopeček, I., Bártek, L., Web Pages for Blind People - Generating Web-Based Presentations by means of Dialogue. In Computers Helping People with Special Needs -P

Abstrakt

Popsaný přístup umožňuje snadno vytvářet grafiku, jejíž popis je čitelně zahrnut do formátu grafického objektu takovým způsobem, že jej lze využívat k automatickému získání popisu vygenerované grafiky na webovských presentacích, které se tak stávají přístupnějšími pro nevidomé uživatele.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
24	Kopeček, Ivan - Bártek, Luděk. Web Pages for Blind People - Generating Web-Based Presentations by means of Dialogue. In Computers Helping People with Special Needs -Proceedings of ICCHP 2006. Berlin - Heidelberg : Springer, 2006,. od s. 114-119, 6 s. ISBN 0302-9743.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/25/2006**

Název výsledku

Hlaváčková, Dana - Horák, Aleš - Kadlec, Vladimír. Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. Lecture Notes in Artifi

Abstrakt

V články se analyzují možnosti, jak využít valenčních rámců z databáze Verbalex v automatické syntaktické analýze češtiny.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
25	Hlaváčková, Dana - Horák, Aleš - Kadlec, Vladimír. Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2006, Berlin, Heidelberg : Springer, 2006, 4188, od s. 79-86, 8 s. ISSN 0302-9743. 2006.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/26/2006**

Název výsledku

Pala, K. Word Sketches and Semantic Roles, Proceedings of the Corpus 2006 Conference, St. Petersburg University, St. Petersburg, 2006, in print.

Abstrakt

V článku se rozebírá využití korpusových dat pro verifikaci valenčních rámců a inventáře sémantických rolí, jež se v rámcích využívají.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
26	Pala, K. Word Sketches and Semantic Roles, Proceedings of the Corpus 2006 Conference, St. Petersburg University, St. Petersburg, 2006, in print.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/27/2006**

Název výsledku

Horák, Aleš - Kadlec, Vladimír. Platform for Full-Syntax Grammar Development Using Meta-grammar Constructs. In Proceedings of the 20th Pacific Asia Conference

Abstrakt

V textu autoři věnují pozornost platformě pro budování gramatik,v níž se využívají metagramatická schémata.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
27	Horák, Aleš - Kadlec, Vladimír. Platform for Full-Syntax Grammar Development Using Meta-grammar Constructs. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. 2006. vyd. Beijing, China : Tsinghua University Press, 2006,. od s. 311-318, 8 s. ISBN 7-302-14060-X. 2C06009.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/28/2006**

Název výsledku

M - Third International WordNet Conference, GWC 2006, Seogwipo, Korea, January 22-26, 2006.

Abstrakt

Sojka, Petr - Choi, Key-Sun - Fellbaum, Christiane - Vossen, Piek. Proceedings of the Third International WordNet Conference, GWC 2006, Seogwipo, Korea, January 22-26, 2006. Edited by Sojka P., Choi K.-S., Fellbaum Ch., Vossen P. prvni. Brno : Masaryk University, 2006,. 362 s. GWC Proceedings. Third International WordNet Conference, GWC 2006, Brno, Czech Republic, January 22--26, 2006, Proceedings. ISBN 80-210-391.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
28	Sojka, Petr - Choi, Key-Sun - Fellbaum, Christiane - Vossen, Piek. Proceedings of the Third International WordNet Conference, GWC 2006, Seogwipo, Korea, January 22-26, 2006. Edited by Sojka P., Choi K.-S., Fellbaum Ch., Vossen P. prvni. Brno : Masaryk University, 2006,. 362 s. GWC Proceedings. Third International WordNet Conference, GWC 2006, Brno, Czech Republic, January 22--26, 2006, Proceedings. ISBN 80-210-391.	M - Uspořádání (zorganizování) konference	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/29/2006**

Název výsledku

M - Mezinárodní konference Text, Speech and Dialogue 2006

Abstrakt

Sojka, P. - Kopeček, I. - Pala, K. Ninth International Conference on TEXT, SPEECH and DIALOGUE (Devátá mezinárodní konference o textu, řeči a dialogu)TSD.2006. Proceedings was published by Springer Verlag as LNAI 4188. http://www.tsdconference.org/tsd2006/

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
29	Sojka, P. - Kopeček, I. - Pala, K. Ninth International Conference on TEXT, SPEECH and DIALOGUE (Devátá mezinárodní konference o textu, řeči a dialogu)TSD.2006. Proceedings was published by Springer Verlag as LNAI 4188. http://www.tsdconference.org/tsd2006/	M - Uspořádání (zorganizování) konference	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/30/2006**

Název výsledku

Bartošek, Miroslav - Lhoták, Martin - Rákosník, Jiří - Sojka, Petr - Šárfy, Martin. DML-CZ: The Objectives and the First Steps. A.K. Peters Ltd., 14 pp., accept

Abstrakt

Článek pojednává o klasifikaci textů - byla shromážděna rozsáhlá sada dokumentů, a to již částečně klasifikovaných dle normy MSC 2000 (Mathematical Subject Classification) a provedeno její předzpracování pro potřeby strojového učení. Jde o články digitalizované v rámci projektu DML CZ

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
30	Bartošek, Miroslav - Lhoták, Martin - Rákosník, Jiří - Sojka, Petr - Šárfy, Martin. DML-CZ: The Objectives and the First Steps. A.K. Peters Ltd., 14 pp., accepted for publication, in print, 2007.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/31/2006**

Název výsledku

Olivia Sanchez, Massimo Poesio, Mijail A. Kabadjov, Roman Tesar: What kind of problems do protein interactions raise for anaphora resolution? A preliminary anal

Abstrakt

In this preliminary study, we analyzed the kind of anaphoric expressions that occur in expressions describing protein interactions found in biological text. We also studied the impact of anaphora resolution on protein interaction extraction, when an off-the-shelf anaphoric resolver (i.e., not one specially developed for this domain) is used, and looking at full texts as well as abstracts. Our results suggest that about 5% of the descriptions of protein-protein interactions contain anaphoric expressions when full texts are considered. These anaphoric expressions are primarily pronouns, even though most anaphoric expressions are full NPs. The use of our anaphoric resolver gives a small improvement over our baseline system.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno

Spojení

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
31	Olivia Sanchez, Massimo Poesio, Mijail A. Kabadjov, Roman Tesar: What kind of problems do protein interactions raise for anaphora resolution? A preliminary analysis. ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-177/	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

4.1.3. PLNĚNÍ DÍLČÍCH CÍLŮ

Dílčí cíl nebyl pro rok 2006 plánován. Příloha "4.1.3. PLNĚNÍ DÍLČÍCH CÍLŮ" se nezpracovává.

4.1.4. REDAKČNĚ UPRAVENÁ ZPRÁVA

Cílem projektu "Prostředky tvorby komplexní báze znalostí pro komunikaci se sémantickým webem v přirozeném jazyce" je vývoj nástrojů pro dokonalejší komunikaci s webovým prostředím. V r. 2006 byly po ustavení řešitelského týmu vytvářeny korpusy pro ověřování algoritmů komunikace, potřebná infrastruktura a testovány nově navržené či modifikované postupy. Dílčí výsledky byly publikovány v 29 článcích.

4.1.5. PLNĚNÍ PODMÍNEK PROGRAMU

Plnění specifických podmínek programu - se pro projekty NPV II nezpracovává. Pro projekty NPVII specifické podmínky ve vyhlášení programu nebyly formulovány.

4.1.6. PLNĚNÍ SMLOUVY O SPOLUPRÁCI

Na základě vymezených základních práv (viz uzavřená smlouva upravující vztahy mezi příjemcem a spolupříjemcem) příjemce poskytnul spolupříjemci finanční dotaci přímým převodem na stanovený účet Masarykovy univerzity, náklady na projekt byly vedeny v oddělené evidenci obou spolupracujících subjektů. Uzavřená smlouva o spolupráci je plněna beze zbytku, plánované finanční prostředky byly vyčerpány - viz odstavec 2.3.2.

4.2. DALŠÍ PŘÍLOHY - rok 2006

4.2.1. Odborné a věcné přílohy zprávy - seznam

	Soubor	
	Zprava_2C06009_odst421.doc	

4.2.2. Ostatní (např. možné využití výsledků) - seznam

	Soubor	
	acm-tesar.pdf	
	CISE2006final.pdf	
	corpora2006vmjmrm.pdf	
	elpub-F117.pdf	
	ESSV_2006.pdf	
	icassp2006.pdf	
	ictta06.pdf	
	inscit06.pdf	
	phdws2006konopik.pdf	

4.2.3. Zápisy z projednání (oponentské posudky) - seznam

	Soubor	
	V elektronické podobě soubor nebyl nositelem poskytnut.	

4.2.4. Zápisy a dokumenty z jednání se styčnými pracovníky zadavatele - seznam

	Soubor	
	V elektronické podobě soubor nebyl nositelem poskytnut.	

4.2.5. Zápisy z jednání Rady projektu (Centra) - seznam

Příloha 4.2.5. Zápisy z jednání Rady projektu (Centra) - se pro tento program nezpracovává.

4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodněním - seznam

Příloha 4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodnění - se pro tento program nezpracovává.